社団法人日本補綴歯科学会　第118回学術大会　研究セミナー 2009.6.4.

多変量解析の使い方のこつ

国立保健医療科学院人材育成部
横山徹爾

多変量解析の分類

目的変数		解析の目的	説明変数	
			量的	質的
あり	量的	関係式の発見 量の推定	重回帰（型の）分析 正準相関分析	数量化分析Ⅰ類
	質的	標本の分類 質の推定	クラスター分析 判別分析	クラスター分析 数量化分析Ⅱ類
なし		変量の整理 変量の分類 代表変量の発見	主成分分析 因子分析 MDS（多次元尺度構成法）	数量化分析Ⅲ，Ⅳ類

多変量解析が難解と思われる最大の理由
＝手法が細分化されているため，どの手法を用いればよいかわかりにくい．

本日扱うのは（　）だけ。

数量化分析Ⅰ類＝ダミー変数を用いた重回帰分析
数量化分析Ⅱ類＝ダミー変数を用いた判別分析
数量化分析Ⅲ類＝ダミー変数を用いた主成分分析
数量化分析Ⅳ類＝ダミー変数を用いたMDS

多変量解析

- 複数の変数からなる多変量データを扱う統計的手法の総称。
- 重回帰分析は、厳密には多変量解析ではないという意見もあるが、医学分野で多変量解析と称して最も使われているのは重回帰分析（型の方法）。
- 因子分析・主成分分析は多変量解析なのは自明の理なので、わざわざ多変量解析と称することは少ない。
- 本日は重回帰分析（型の方法）を中心に。

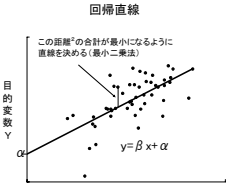
本日の予定

- ①重回帰分析（型の方法）
  - 重回帰分析
    - その特殊型：共分散分析、偏相関分析
  - 多重ロジスティックモデル
  - 多変量Cox比例ハザードモデル
    - 類似のモデル：多変量ポアッソン回帰
- ②主成分分析・因子分析

多変量解析（重回帰分析）を始める前に

- いきなり多変量解析をしない！
  - ヒストグラム、平均、標準偏差等でデータの特徴を把握。
    - 多くの統計手法は正規分布を仮定しているので、必要に応じて対数変換等を考慮。
  - （単）相関分析により、多数の変数間の相関の強さを確認する。
    - 相関の強い2つ以上の変数を同時に説明変数に含めると、変な結果になることがあるので注意。
  - （単）回帰分析を行った後、重回帰分析に進み、結果がどう変わるかをよく見る。
    - （単）回帰分析をせずに重回帰分析をすると、解釈困難。

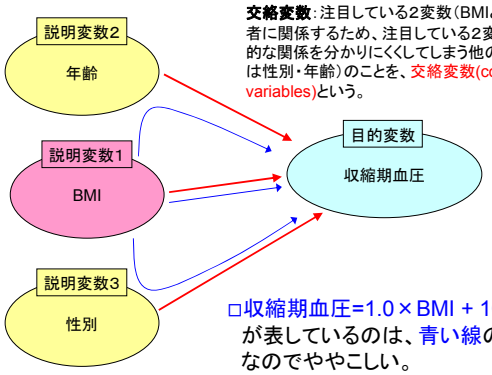
（単）回帰分析　regression analysis



- 2つの測定値y（目的変数）とx（説明変数）の関係を、$y=\beta x+\alpha$ の形の1次式（回帰式）で表す。
- 回帰係数 β（傾きともいう）
  - 説明変数xが1増加した時の、目的変数yの増加量の期待値を表す。
- 切片 α
  - x=0のときのyの期待値。
- yは連続量で正規分布。xは連続量、2値データなど。
- 検定
  - 「帰無仮説 H_0: $\beta=0$」を検定し有意ならば、xとyは有意な直線的関係があると解釈。
- 例
  - 男性100人の収縮期血圧を目的変数y、BMIを説明変数xとする回帰分析を行ったところ、回帰式はy=1.0 x + 100と推定され、帰無仮説：母回帰係数=0は有意水準5%で棄却された。
- 解釈
  - BMIが1大きいと収縮期血圧は1mmHg高いことが期待されるという有意な直線的関係がある。

回帰分析がたくさん

- ある1つの目的変数yと、3つの説明変数x_1～x_3の関係を、3つの回帰式で表すことを考える。
  - $y=\beta_1 x_1+\alpha_1$　例）収縮期血圧=1.0×BMI + 100
  - $y=\beta_2 x_2+\alpha_2$　例）収縮期血圧=0.5×年齢 + 110
  - $y=\beta_3 x_3+\alpha_3$　例）収縮期血圧=15.0×性別(男1,女0) + 120
- 解釈
  - BMIが1大きいと収縮期血圧は1mmHg高い
  - 年齢が1歳大きいと収縮期血圧は0.5mmHg高い
  - 男性は女性に比べて収縮期血圧は15.0mmHg高い
  - BMIが25.0の人の収縮期血圧は125mmHgと予測される･･･など。
  - まずは、おおざっぱな傾向を把握してから次に進む。
- 3つを同時に解釈できるか？
  - BMIが大きい人は、高齢者が多く、男性が多いかも。
  - すると、「BMIが1大きいと収縮期血圧は1mmHg高い」は本当は年齢や性別の影響なのでは？
  - 同様に、「年齢が1歳大きいと収縮期血圧は0.5mmHg高い」は本当は年齢ではなくてBMIのせいなのでは？
  - BMIが25で60歳で男性の収縮期血圧はいくつ？125mmHg?140mmHg?135mmHg?
  - など･･･解釈困難である。



交絡変数：注目している2変数（BMIと血圧）の両者に関係するため、注目している2変数間の直接的な関係を分かりにくくしてしまう他の変数（ここでは性別・年齢）のことを、交絡変数(confounding variables)という。

□収縮期血圧=1.0×BMI + 100が表しているのは、青い線の関係なのでややこしい。

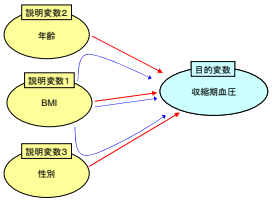
□知りたいのは赤い線の関係。そこで･･･重回帰分析。

観察研究では交絡変数の調整は必須

- 無作為化比較試験では、群間で交絡変数の分布がずれることは少ない（あるとしても偶然の範囲）。
  - □そのため、多変量解析を用いる必要性は乏しい（主解析にはあまり用いない）。
- 観察研究では、交絡変数がほぼ必ず存在。
  - □従って、主要な解析結果は交絡変数の影響を調整した多変量解析とすべき。

重回帰分析 multiple linear regression analysis

- ある1つの目的変数yと、3つの説明変数x_1～x_3の関係を、1つの1次式（重回帰式）で表すことを考える。
  - $y=\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\alpha$　（β_1～β_3：偏回帰係数）
  - 例）収縮期血圧=0.9×BMI + 0.4×年齢 + 16×性別(男1,女0) + 95
- 解釈
  - 年齢と性別とは独立に（年齢と性別の影響を除いても・調整しても）、BMIが1大きいと収縮期血圧は0.9mmHg高い
  - 性別とBMIとは独立に（性別とBMIの影響を除いても・調整しても）、年齢が1歳大きいと収縮期血圧は0.4mmHg高い
  - 年齢とBMIとは独立に（年齢とBMIの影響を除いても・調整しても）、男性は女性に比べて収縮期血圧は16.0mmHg高い
  - BMI=25, 年齢60歳, 男性の収縮期血圧は0.9×25+0.4×60+16×1+95=157.5mmHgと予測される。



（単）回帰分析と重回帰分析の比較

	収縮期血圧					
	（単）回帰分析			重回帰分析		
	回帰係数	標準誤差	P値	偏回帰係数	標準誤差	P値
飲酒量（合）	4.0	0.5	<0.001	4.1	0.5	<0.001
喫煙量（箱）	2.0	0.9	0.02	0.5	0.8	0.90

- （単）回帰分析の解釈
  - 飲酒量が1合多いと、血圧は4mmHg高いが、これに含まれる喫煙の影響はわからない。
  - 喫煙量が1箱多いと、血圧は2mmHg高いが、これに含まれる飲酒の影響はわからない。
- 重回帰分析の解釈
  - 喫煙とは独立に（の影響を除いても・調整しても）、飲酒量が1合多いと、血圧は4.1 mmHg高い。
  - 飲酒とは独立に（の影響を除くと・調整すると）、喫煙量と血圧の関係は明らかでない。

- 他の説明変数の影響を調整したうえで、目的変数と説明変数間の関連を調べるのが重回帰分析。
- 同時に用いた説明変数によって、解釈が少し変わる。

標準化偏回帰係数

- BMIと年齢のどちらの方が収縮期血圧との関連が強いか考えたいとき、偏回帰係数を比較しても無意味（なぜなら、単位が異なるから）
- そこで、全ての変数を平均=0、標準偏差=1となるようにに変換（標準化）してから重回帰分析を行い得られた偏回帰係数のことを、標準化偏回帰係数という。
- どの説明変数との関連が強いのかを解釈しやすい。

寄与率R^2

- 重回帰分析を行い、回帰式
  - 収縮期血圧=0.9×BMI + 0.4×年齢 + 16×性別(男1,女0) + 95
- を作った際、寄与率R^2（決定係数）を計算することがある。
  - 例）R^2=0.30
  - 収縮期血圧の個人差（分散）のうち、30%をBMIと年齢と性別によって説明できるということを意味する。
- また、BMI, 年齢, 性別のそれぞれの説明変数についても、偏寄与率(partial R^2)を計算できる。

重回帰分析による収縮期血圧の関連要因

	偏回帰係数	標準誤差	P値	偏R^2
BMI	0.9	0.3	0.003	0.08
年齢	0.4	0.2	0.046	0.13
性別	16.0	1.2	<0.001	0.12

モデルR^2=0.30

- 解釈
  - BMIと年齢と性別によって、収縮期血圧の全分散（個人差）の30%が説明できる。
  - 個々の要因で見ると、BMI単独で8%、年齢単独で13%、性別単独で12%である。（合計は必ずしも30%にならないので注意）
- R^2が小さいと、これらの説明変数では収縮期血圧の個人差を十分に説明できない（十分に予測できない）という意味。
  - R^2が小さいと、モデルに「意味がない」と極端な言い方をする人がいるが、「予測にはあまり役立たない」と言う方が適切。R^2が小さくても関連を調べるだけならば、十分に意味のある分析に成り得る。

重回帰分析の説明変数に関する注意

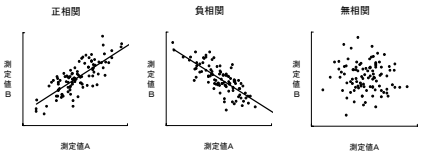
- 全く同じ意味を持つ2変数を同時に使ってはいけない。
  - 例）2回測定した血圧を、2つとも同時に説明変数に入れるのはナンセンス！
    - 1回目の血圧で調整した2回目の血圧って・・・ただのノイズ？
- 類似の理由で、相関が非常に強い2変数を同時に使うのは、望ましくないことが多い。
  - 「多重共線性」の問題が生じ、推定値が極めて不安定になることがある。特に、単回帰と重回帰の結果が大きく変わった時（回帰係数の符号が変わったとか）は要注意。
- 変数のもつ医学的な意味が変わることがあるので注意。
  - 例1）収縮期血圧SBPと拡張期血圧DBPを同時に入れると、DBPで調整したSBPって・・・“脈圧みたいなもの”？
  - 例2）身長と体重を同時に入れると、身長で調整した体重って・・・“肥満度みたいなもの”？

重回帰分析における変数選択

- 重回帰分析では、多数の説明変数を同時に扱うことができるが、前記の理由により、何でもかんでも入れれば良いというものではない。変数の数が多すぎると、解釈が困難だったり、推定値が不安定になることがある。
- まず、研究仮説と医学的な意味をよく考えて投入する変数を吟味する。
- 様々な説明変数の組合せのうち、有意なものだけを選び出すために、stepwise法などによる変数選択を行う。
  - その際、単回帰分析で有意なものだけを候補にすることもある。
  - 決定係数R^2が大きいほど予測性能が高いことを意味するので、全ての組合せの中からR^2最大のものを選ぶという方法もあるが、計算が大変等の理由からあまり使われない。
- Stepwise法
  - 変数を一つずつ追加していく（変数増加法）
  - 一つずつ減らしていく（変数減少法）
  - 有意でないものを除去し、有意になるものを投入するという繰り返しで選んでいく（変数増減法）
    - 必ず調整すべき変数は強制的に含めることができる。

相関分析 correlation analysis

図7 正相関と負相関



- 回帰分析と非常によく似ている。
  - 回帰分析はXとYに単位があるが、相関分析にはない。
  - 回帰分析は目的変数と説明変数の区別があるが、相関分析にはない。
- 相関係数 correlation coefficient
  - -1～+1の値をとり、2変数の直線的な関連の強さを表す。
  - 単位がないので、様々な変数間で関連の強さを比較するのに便利。（単位がある回帰分析では、kgとgの違いだけで回帰係数が1000倍変わってしまうが、相関係数は変わらない）
  - 検定も行う（帰無仮説：母相関係数=0）。検定結果は回帰分析のものと一致する。
  - 正規分布する2つの連続量の場合に用いる（Pearsonの（積率）相関係数ともいう）。
  - 外れ値があると変な値をとることがあるので、正規分布でない変数の場合には、Spearmanの順位相関係数を用いることが多い。

偏相関分析 partial correlation analysis

- 重回帰分析と非常によく似ている。
  - 例）収縮期血圧、BMI、年齢、性別それぞれの独立な関係を調べて、偏相関係数で表す。
- 偏相関係数 partial correlation coefficient
  - -1～+1の値をとり、他の変数とは独立な（影響を除いた・調整した）2変数の直線的な関連の強さを表す。
  - 重回帰分析の偏寄与率の平方根に符号を付けたものである。
  - 検定も行う（帰無仮説：母偏相関係数=0）。検定結果は重回帰分析のものと一致する。

重回帰分析とダミー変数

- $y=\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\alpha$
  - 説明変数xは一般に、連続量や2値変数。
  - xに名義尺度を使いたいときはどうする？
    - 例）喫煙の、①吸う、②やめた、③吸わない
    - ③吸わない人に比べて、①吸う人、②やめた人は、それぞれ血圧はどのくらい異なるのだろうか？
  - そこで、以下のように変数（ダミー変数）を作り、
- $y=\beta_1 x_1+\beta_2 x_2+\alpha$
  - という重回帰分析を行う。

①吸うvs.③吸わない　②やめたvs.③吸わない

	ダミー変数		
	x_1	x_2	x_3
①吸う人	1	0	0
②やめた人	0	1	0
③吸わない人	0	0	1

重回帰分析とダミー変数（続き）

- $y=\beta_1 x_1+\beta_2 x_2+\alpha$

①吸うvs.③吸わない　②やめたvs.③吸わない

	ダミー変数		
	x_1	x_2	x_3
①吸う人	1	0	0
②やめた人	0	1	0
③吸わない人	0	0	1

		偏回帰係数	標準誤差	P値
①吸う人	(β_1)	3.5	1.3	0.007
②やめた人	(β_2)	5.4	2.2	0.014
③吸わない人		基準	–	

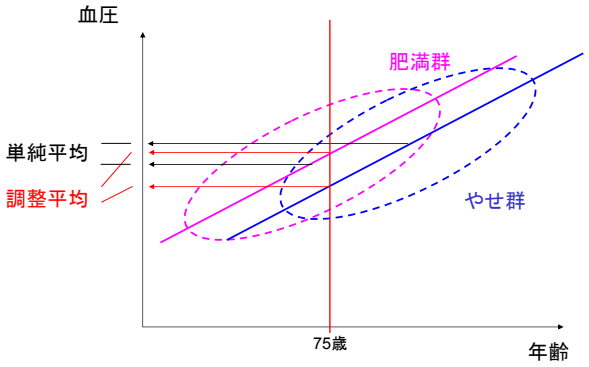
- 解釈
  - 吸わない人に比べて、吸う人は3.5mmHg、やめた人は5.5mmHg、有意に血圧が高かった。
- ダミー変数を用いた重回帰分析は、数量化Ⅰ類と本質的に同じ。

共分散分析と調整平均

	収縮期血圧		年齢
	平均	標準偏差	平均
BMI			
18.5以下	130	15	85
18.5以上25未満	125	12	70
25以上	122	12	65

- やせている人ほど血圧が高い？
  - 実は、やせている人ほどお年寄りが多かった。
  - こんなとき、年齢の影響を調整した平均値が計算できると便利。
- 共分散分析（重回帰分析の特殊型）により、調整平均（調整最小2乗平均）を計算可能。

共分散分析ANCOVA(Analysis of Covariance)による調整平均（最小2乗平均LSM: Least Square Mean）



共分散分析と調整平均

	収縮期血圧		年齢
	年齢調整平均※	標準誤差	平均
BMI			
18.5以下	123	1.5	85
18.5以上25未満	126	1.3	70
25以上	128	1.8	65

※共分散分析による年齢調整最小2乗平均。
P<0.05 for trend.

- このように示せば、年齢の影響を調整すると太っている人の方が血圧が高いことがよくわかる。

重回帰分析で、2元配置分散分析と同じことをする（1）

歯科材料への着色の程度（値は平均±SD）

		フッ素	
		（－）	（＋）
紅茶色素	（－）	1.0±0.4	2.0±0.5
	（＋）	3.0±0.6	4.0±0.5

紅茶色素の効果 β_2=2.0, P=0.01

フッ素の効果 β_1=1.0, P=0.05

- 二元配置分散分析
  - アウトカム（着色）に及ぼす、二つの要因（フッ素、紅茶色素）の独立な影響を分析する。
    - フッ素の影響と、紅茶の影響と、分離して評価できる。
- 重回帰分析
  - $y=\beta_1 x_1+\beta_2 x_2+\alpha$
    - x_1: フッ素(-)=0, (+)=1, x_2: 紅茶色素(-)=0, (+)=1

重回帰分析で、2元配置分散分析と同じことをする（2）

歯科材料への着色の程度（値は平均±SD）

		フッ素	
		（－）	（＋）
紅茶色素	（－）	1.0±0.4	2.0±0.5
	（＋）	3.0±0.6	7.0±0.5

紅茶色素の効果=??

フッ素の効果=??

- フッ素の有無によって紅茶色素の効果が変わる（逆も同様）。
- そのため、紅茶色素の効果とフッ素の効果を単純には示せない。
- 交互作用という概念が必要。

歯科材料への着色の程度（値は平均±SD）

		フッ素	
		（－）	（＋）
紅茶色素	（－）	1.0±0.4	2.0±0.5
	（＋）	3.0±0.6	7.0±0.5

紅茶色素の主効果=2.0 β_2

フッ素の主効果=1.0 β_1

交互作用=3.0 β_{12}
P=0.03

1.0+2.0+1.0=4.0のはずのところが7.0になっているので

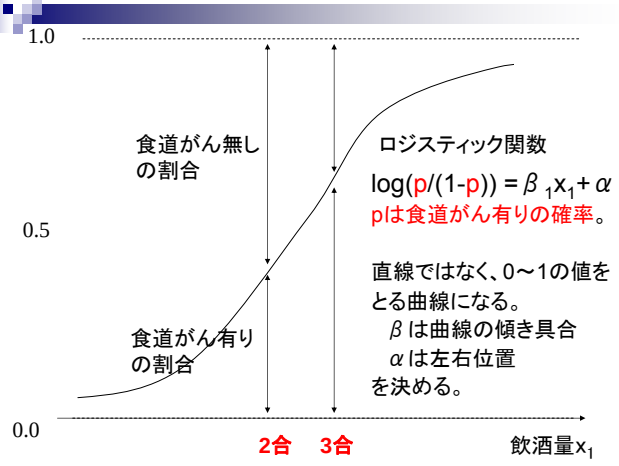
- 二元配置分散分析（交互作用あり）
- 重回帰分析（交互作用あり）
  - $y=\beta_1x_1+\beta_2x_2+\beta_{12}x_1x_2+\alpha$
    - x_1: フッ素(-)=0, (+)=1, x_2: 紅茶色素(-)=0, (+)=1
  - フッ素と紅茶色素が単独の時の効果がそれぞれの主効果。
  - 同時に組み合わさった時に、主効果の和にさらに上積みされる効果が交互作用。“フッ素×紅茶色素”のようにかけ算する。
  - 交互作用がある時は、主効果だけでの解釈はしない。交互作用も見て、総合的に解釈する。

本日の予定

- ①重回帰分析（型の方法）
  - 重回帰分析
    - その特殊型：共分散分析、偏相関分析
  - 多重ロジスティックモデル
  - 多変量Cox比例ハザードモデル
    - 類似のモデル：多変量ポアッソン回帰
- ②主成分分析・因子分析

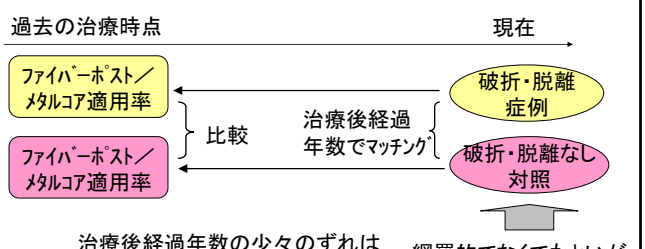
重回帰分析型の他の方法（1）

- 重回帰分析
  - $y=\beta_1x_1+\beta_2x_2+\beta_3x_3+\alpha$
  - yは正規分布する連続量
  - yが病気の有り(=1)、なし(=0)の2値の場合には使えない。
    - 正規分布の仮定に反するだけでなく、xの値によってはyの予測値が1を超えたり0を下回ったりすることがある→解釈不能。
- 多重ロジスティックモデル
  - $\log(p/(1-p))=\beta_1x_1+\beta_2x_2+\beta_3x_3+\alpha$
  - pは病気有りの確率。p/(1-p)を病気ありのオッズ、log(p/(1-p))をpのロジットという。
  - 目的変数が病気の有り、なし、のように2値の場合によく使われる方法。
  - exp(β)によって、オッズ比を計算できる。
  - オッズ比は、説明変数x（例えば喫煙=1, 非喫煙=0）があると、病気ありのオッズ（稀な疾病では≒確率）が何倍になるかを意味する指標である。
  - 2×N分割表を、交絡変数で調整して検定するという目的にも使える。



多重ロジスティックモデルと症例・対照研究

あるイベントを起こした症例と、そうでない対照とで、それと関連する要因を過去にさかのぼって調べる研究方法。



治療後経過年数の少々のずれは統計学的方法で調整可能

網羅的でなくてもよいが偏りなく選ぶ必要あり（対照の選び方が難しい）

解析1　基本属性の比較

	破折・脱離症例 n=90	対照 n=90	P値
治療後経過年数	5.0±2.1年	5.1±2.2年	マッチング
ファイバーポスト%	13%	25%	<0.001
男性割合	70%	65%	0.04
年齢	55.1±7.2	50.2±10.8	0.01
セメント（レジン%）	48%	46%	0.88
ポストの長さ（対歯根長比）	0.55±0.05	0.50±0.05	0.02

値は平均±標準偏差

マッチングしているので、対応のある検定を行う（Wilcoxon符号付き順位検定、McNemar検定）

解析2　オッズ比

		破折・脱離症例		
		ファイバーポスト	メタルコア	計
対照	ファイバーポスト	2	21	23
	メタルコア	10	57	67
	計	12	78	90

値はペア数

症例対照研究では、オッズ比≒相対危険度
オッズ比（マッチングした場合）= 10 / 21 = 0.48
多変量解析には、条件付きロジスティック回帰を用いる。

解析3　多変量調整オッズ比

	破折＋脱離 ｵｯｽﾞ比(95%信頼区間)
ﾌｧｲﾊﾞｰﾎﾟｽﾄvs.ﾒﾀﾙｺｱ	0.55 (0.45−0.67)
男性　vs. 女性	1.18 (0.97−1.44)
年齢　+10歳あたり	1.02 (0.83−1.24)
セメント（レジンvs.他）	0.88 (0.72−1.07)
ポストの長さ　+1SDあたり	1.15 (0.94−1.40)

解釈：これらの交絡変数で調整しても、ﾌｧｲﾊﾞｰﾎﾟｽﾄは破折、脱離のリスクが低い。（未知の交絡の影響は不明）

値は仮想データです

ここでは全ての変数を同時に考慮したが、特定のアルゴリズムで有意なものだけを選ぶことも多い（stepwise法など）

クロス表の検定への応用

喫煙状況と体重との関係

	やせ	普通体重	肥満	平均年齢
喫煙（n=200）	20%	50%	30%	60
非喫煙（n=300）	10%	70%	20%	70

年齢調整P=0.01

- このような2×3分割表でよく使うのが、χ^2検定。
- しかし、喫煙群と非喫煙群で平均年齢が異なっていると、喫煙の影響なのか年齢の影響なのか分からない。
- そこで、喫煙の有無を目的変数、体重（ダミー変数）と年齢を説明変数にした多重ロジスティックモデルを用いることで、年齢調整したうえで喫煙状況と体重との関係を検定することができる。
  - $\log(p/(1-p)) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \alpha$
  - pは喫煙ありの割合、x_1, x_2は体重を表すダミー変数、x_3は年齢。
- Cochran-Mantel-Haenszel法でも可

判別を目的とした応用

- 多重ロジスティックモデルは、
  - $\log(p/(1-p)) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \alpha$
  - pは病気有りの確率
- だから、説明変数を与えると、ある人が病気である“確率”を計算することができる。これを、疾病の有無のような判別目的に応用することができる。
- 古典的な判別分析は、ある人の所属する群を1つに決めようとするが、多重ロジスティックモデルでは所属する群を“確率”で表現できる。
- 例）ある複数の検査値から、その人が疾病Aであるか否かを判別したい。
  - 古典的な判別分析：　Aであるか否かのどちらかに分類。
  - ロジスティックモデル：　Aである確率が○○%のように予測。

表1. 日本人男性を対象とした症例・対照研究（234症例、634対照）による、アルコール・フラッシング反応、飲酒、喫煙、緑黄色野菜・果物摂取と、食道扁平上皮癌リスク

危険因子			食道扁平上皮癌の多変量調整オッズ比[a] オッズ比	95%信頼区間
フラッシング反応	飲酒量			
任意		ほとんど飲まない	1	(基準)
なし		少量	1.27	0.27 - 5.88
		中等量	10.12	3.45 - 29.69
		多量	15.61	5.19 - 46.91
		やめた	27.31	5.24 - 142.46
あり		少量	6.69	2.21 - 20.20
		中等量	42.66	14.17 - 128.42
		多量	72.86	23.75 - 223.57
		やめた	37.00	7.66 - 178.76
強いアルコール飲料をよく飲む[c]			3.59	1.63 - 7.87
喫煙30パック年以上[d]			2.62	1.71 - 4.00
緑黄色野菜を毎日は食べない[e]			1.65	1.03 - 2.64
果物を毎日は食べない[e]			1.57	0.94 - 2.62

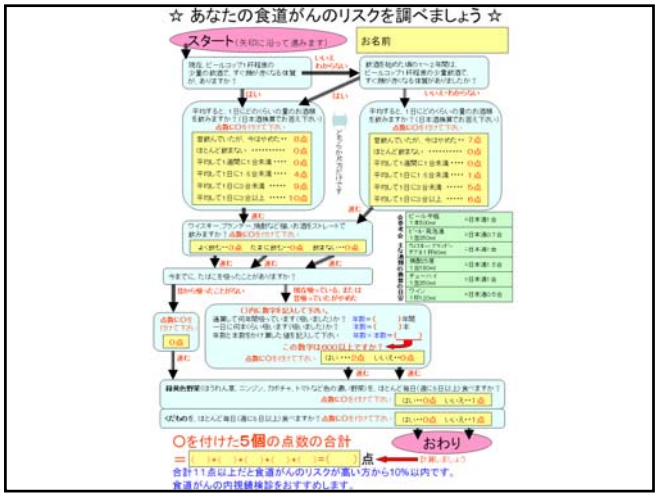
[a] 多重ロジスティックモデルにより全ての変数を同時に調整したオッズ比（年齢も調整したが示していない）。Yokoyama T, et al. (2003) *Cancer Epidemiol Biomarkers Prev*.
[b] ほとんど飲まない：1合未満/週、少量：1～8.9合/週、中等量：9～17.9合/週、多量：18合以上/週.
[c] 「よく飲む」対「飲まない／ときどき」（基準）
[d] 「30パック年以上」対「未満」（基準）
[e] 「毎日以外」対「毎日」（基準）

図2. 食道癌リスク評価のための健康危険度評価(HRA)モデル[アルコール・フラッシングを用いるHRA-Fモデル]。A～Eの合計を、合計リスク得点とする。得点が高いほど食道癌リスクが高いことを意味する。例えば、合計リスク得点が4.57以上の男性の食道癌リスクは、この集団において危険な方から上位10%に相当する。

危険因子		得点（A-Eについてそれぞれ1つずつ選ぶ） オリジナル得点	（整数化得点）	
アルコール・フラッシングと飲酒量				
フラッシング任意				
ほとんど飲まない	(<1合/週)	0.00	(0)	
フラッシングなし				
少量	(1-8.9合/週)	0.24	(1)	
中等量	(9-17.9合/週)	2.31	(5)	
多量	(18合以上/週)	2.75	(6)	
やめた		3.31	(7)	
フラッシングあり				
少量	(1-8.9合/週)	1.90	(4)	
中等量	(9-17.9合/週)	3.75	(9)	
多量	(18合以上/週)	4.29	(10)	A
やめた		3.61	(8)	
強いアルコール飲料をよく飲む				
	はい	1.28	(3)	B
	いいえ	0.00	(0)	
喫煙30パック年以上				
	はい	0.96	(2)	C
	いいえ	0.00	(0)	
緑黄色野菜を毎日食べる				
	はい	0.00	(0)	D
	いいえ	0.50	(1)	
果物を毎日食べる				
	はい	0.00	(0)	E
	いいえ	0.45	(1)	

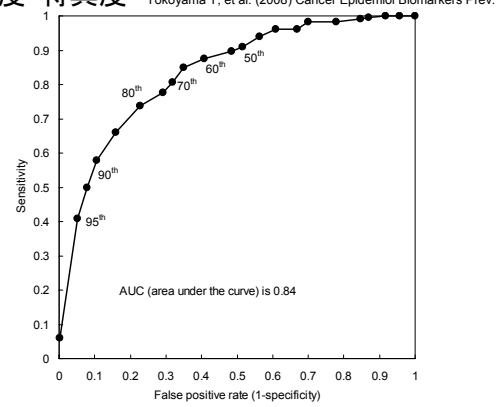
合計得点 = A + B + C + D + E

予測リスク	合計得点 オリジナル	整数化
下位25%	≤1.18	0-2
25-49%	1.19-2.78	3-5
50-74%	2.79-3.80	6-8
75-89%	3.81-4.70	9-10
上位10%	4.71+	11+



HRA質問票を用いた食道がんスクリーニングの感度・特異度

Yokoyama T, et al. (2008) Cancer Epidemiol Biomarkers Prev.



本日の予定

- ①重回帰分析（型の方法）
  - 重回帰分析
    - その特殊型：共分散分析、偏相関分析
  - 多重ロジスティックモデル
  - 多変量Cox比例ハザードモデル
    - 類似のモデル：多変量ポアッソン回帰
- ②主成分分析・因子分析

重回帰分析型の他の方法（2）

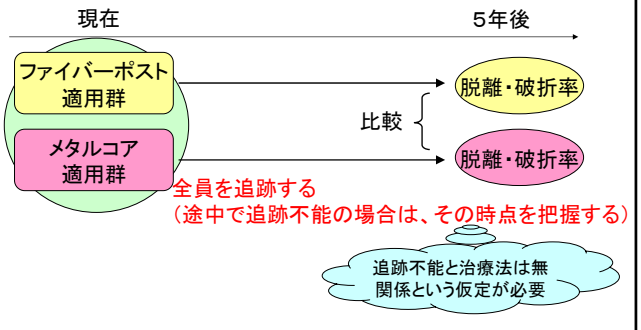
- 多変量Cox比例ハザードモデル
  - $\lambda(t, X) = \lambda_0(t) \exp(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)$
  - $\lambda(t, X)$は、時刻tにおける死亡確率（ハザード）。
  - 目的変数が生存時間の場合によく使われる方法。
  - $\exp(\beta)$によって、相対危険度を計算できる。
  - 相対危険度は、説明変数（例えば喫煙=1, 非喫煙=0）xがあると、死亡確率が何倍になるかを意味する指標である。

コホート研究（仮想例）

ファイバーポストもしくはメタルコアで支台築造した単根歯の生存率比較

- 治療法（無作為割り付け困難）
  - ファイバーポスト
  - メタルコア
- 評価項目
  - 歯の生存率（脱離率、破折率）
  - ファイバーポスト vs. メタルコアで、累積生存率を比較・・・生存時間分析。
- 交絡変数
  - セメント種類、ポストの長さ、経過年数、等
  - 層別分析、多変量Cox比例ハザードモデルによる調整相対危険度・調整生存率

前向きコホート研究



丹後俊郎:「新版・医学への統計学」より

10.1　生存率曲線

まず，生存率曲線（survival curve）とは何かを考えてみよう．次のデータはある発癌物質を投与された7匹のラットの投与時点からの生存日数である．

2，3，3，4，~~7~~，10，11

簡単のため，打切られたデータのない完全データだけをとりあげた．ここで，2日というのは，投与時点からちょうど2日(48時間)経過した時点としよう．この集団の生存率曲線を描けと言われれば，直感的には図71に示す階段曲線となることが理解できるであろうか．

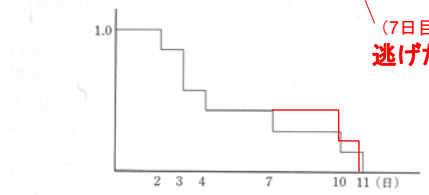


図71　生存率曲線

（7日目に）
逃げた！

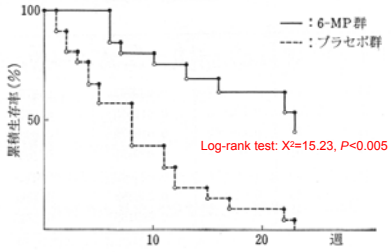


図76　Freireich らによる急性白血病臨床試験における 6-MP 治療群とプラセボ群の生存曲線

Log-rank検定

$$\frac{(\text{A群の観測死亡数}-\text{A群の期待死亡数})^2}{\text{A群の期待死亡数}}+\frac{(\text{B群の観測死亡数}-\text{B群の期待死亡数})^2}{\text{B群の期待死亡数}}\sim\chi_1^2$$

観察研究では、交絡変数の影響が大きいのでこれでは不十分
→多変量Cox比例ハザードモデル

相対危険度

- A群の死亡率Pa、B群の死亡率Pb
- A群に対するB群の死亡の相対危険度
  RR=Pb／Pa
- つまり、死亡率が何倍なのかを表す。
- 死亡率の定義の仕方によって、いろいろな相対危険度の計算方法がある。
  - 累積死亡率を用いる
  - 人・時法死亡率を用いる
  - 瞬間死亡率（＝ハザード）を用いる
    - Cox比例ハザードモデルでは、ハザード比を推定可能。
    - 「比例ハザード性の仮定」の妥当性の検討が必要
      - 参考：丹後俊郎他．ロジスティック回帰分析．朝倉書店(1996).

Cox 比例ハザードモデルの簡単な考え方

正常血圧群と高血圧群を長期間追跡した場合の、死亡の相対危険を考える。

始めに観察集団ありき
　その、１年後、正常血圧群の 1.0%、高血圧群の 2.0%が死亡。……相対危険＝2.0
生き残った人のうち
　さらに１年後、正常血圧群の 1.1%、高血圧群の 2.4%が死亡。……相対危険＝2.2
生き残った人のうち
　さらに１年後、正常血圧群の 1.3%、高血圧群の 2.1%が死亡。……相対危険＝1.6
生き残った人のうち
　さらに１年後、正常血圧群の 1.5%、高血圧群の 3.5%が死亡。……相対危険＝2.3
生き残った人のうち
　さらに１年後、正常血圧群の 1.6%、高血圧群の 3.0%が死亡。……相対危険＝1.9
生き残った人のうち
　さらに・・・・・・・・・・
そして誰もいなくなった。

平均して考えてみると……相対危険は２くらいだった。
（ハザード比=2）

解析1　ベースラインの比較

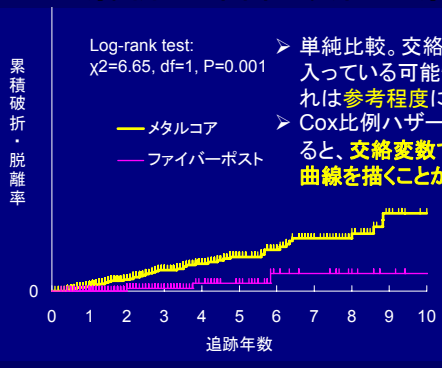
	ファイバーポスト	メタルコア	
	n=200	n=600	P値
男性割合	65%	73%	0.005
年齢	55.1±7.2	50.2±10.8	0.001
セメント（レジン%）	48%	49%	0.52
ポストの長さ（対歯根長比）	0.51±0.05	0.50±0.04	0.65

値は平均±標準偏差

重要な交絡変数にどの程度の違いがあるかを確認する。
必要なものは統計学的に調整する。

値は仮想データです

解析2　評価項目の解析



- 単純比較。交絡変数の影響が入っている可能性があるので、これは参考程度に見ておく。
- Cox比例ハザードモデルを用いると、**交絡変数で調整した生存率曲線を描くことが出来る**。

値は仮想データです

評価項目の解析

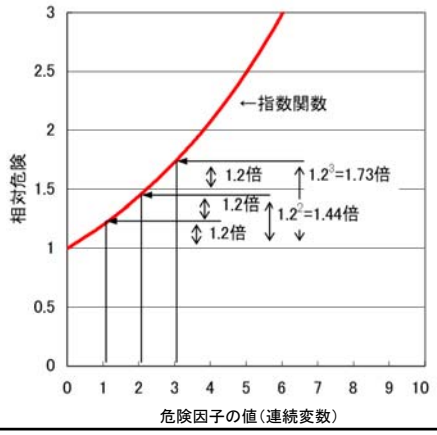
	破折 ﾊｻﾞｰﾄﾞ比(95%信頼区間)	破折＋脱離 ﾊｻﾞｰﾄﾞ比(95%信頼区間)
ﾌｧｲﾊﾞｰﾎﾟｽﾄvs.ﾒﾀﾙｺｱ	0.50 (0.30–0.82)	0.60 (0.49–0.73)
男性　vs. 女性	1.15 (0.70–1.90)	1.20 (1.02–1.41)
年齢　+10歳あたり	1.02 (0.62–1.68)	1.02 (0.97–1.07)
セメント(レジンvs.他)	0.98 (0.59–1.62)	0.90 (0.74–1.10)
ポストの長さ　+1SDあたり	1.30 (0.79–2.14)	1.40 (1.15–1.71)

解釈：これらの交絡変数で調整しても、ﾌｧｲﾊﾞｰﾎﾟｽﾄは破折、脱離のリスクが低い。(未知の交絡の影響は不明)

値は仮想データです

ここでは全ての変数を同時に考慮したが、特定のアルゴリズムで有意なものだけを選ぶことも多い(stepwise法など)

説明変数が1増加した時(1増加あたり)の相対危険度=1.2とは？



本日の予定

- ①重回帰分析(型の方法)
  - 重回帰分析
    - その特殊型：共分散分析、偏相関分析
  - 多重ロジスティックモデル
  - 多変量Cox比例ハザードモデル
    - 類似のモデル：多変量ポアッソン回帰
- ②主成分分析・因子分析

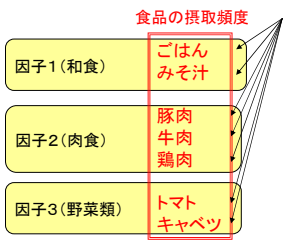
主成分分析と因子分析

- どちらも、多数の変数を少数の変数に要約して表現する方法。
- 主成分分析
  - 観測された多数の変数を、少数の変数に合成して要約する(総合得点化)
- 因子分析
  - 観測された多数の変数から、背後に潜んでいる少数の概念を抽出する(観測された変数の方が合成された変数という点で主成分分析と逆)。

因子分析　factor analysis

- 多数の変数x_1～x_nの背後には、いくつかの概念f_1～f_m(因子という)が潜在しているはずだ。その因子を見つけ出そう！
- $x_1=a_{11}f_1+a_{12}f_2+\ldots+a_{1m}f_m+d_1u_1$
- $x_2=a_{21}f_1+a_{22}f_2+\ldots+a_{2m}f_m+d_2u_2$
- ….

観測されたxはfから合成されている。
aは因子負荷量。



- 因子分析によって、
  - 背後に潜んでいる因子1～3を見つけ出し、数値(因子得点)で表すことができる。
  - 因子1は質問1，3とよく相関する(相関の程度＝因子負荷量)
  - 因子を同定しやすいように、Varimax回転やPromax回転という処理を行うことが多い。

食品摂取習慣と背景因子

食品	因子番号と解釈 1.和食	2.肉食	3.野菜果物
日本茶	**0.82**	-0.12	-0.01
コーヒー	0.08	0.32	0.05
ごはん	**0.73**	0.10	-0.06
パン	**-0.42**	0.24	0.02
めん類	0.11	0.10	-0.04
緑黄色野菜	0.44	0.10	**0.76**
その他の野菜	0.32	0.06	**0.78**
魚料理	**0.67**	**-0.42**	0.13
鶏肉料理	0.31	**0.78**	0.06
牛・豚肉料理	-0.33	**0.84**	-0.01
ハム・ソーセージ	-0.15	**0.54**	-0.10
かんきつ類	0.18	0.03	**0.64**
その他の果物	0.06	-0.04	**0.56**
洋菓子	**-0.42**	0.21	0.02
和菓子	**0.55**	-0.03	0.14
固有値	3.13	2.03	1.90
寄与率	22.4%	14.5%	13.6%

- 因子負荷量
  - 各因子と各食品との相関。
- 固有値
  - 何個分の食品項目の情報を要約しているか。
- 寄与率
  - 全体の分散の何%を要約しているか。

値は仮想データです

値は仮想データです

食習慣と虚血性心疾患罹患リスク

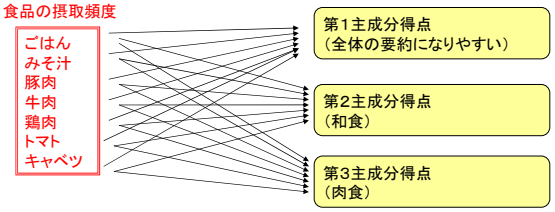
	虚血性心疾患罹患	
	相対危険度	95%信頼区間
第1因子得点(和食)		
低	1(基準)	
中	0.72	0.12-1.07
高	0.51	0.09-0.75
第2因子得点(肉食)		
低	1(基準)	
中	1.32	0.23-1.95
高	2.65	0.46-3.92
第3因子得点(野菜果物)		
低	1(基準)	
中	0.82	0.14-1.21
高	0.58	0.10-0.86

- 因子得点は個人ごとに計算される。
- 他の量的変数と同じように分析に使用できる。
- 個々の食品ではわからなかった特徴が見えることも。

主成分分析 principal component analysis

- 多数の変数x_1～x_nを、いくつかの要約指標z_1～z_n(主成分)にまとめよう。
- $z_1=w_{11}x_1+w_{12}x_2+\ldots+w_{1n}x_n$
- $z_2=w_{21}x_1+w_{22}x_2+\ldots+w_{2n}x_n$
- ....

観測されたxからzを合成する。



国立保健医療科学院における生物統計関連の教育

- 遠隔教育・生物統計学
  - いわゆるe-learning。埼玉県まで来なくても自宅等で受講できる。3ヶ月かけて教科書を1冊学習。
  - 定員30名。
- 臨床試験に係わる臨床医向け生物統計学研修
  - 臨床試験のプロトコルを自分で作って実施しようという臨床医向け。臨床試験に特化した研修で、統計学そのものは時間をあまりかけない。
- 専門課程・生物統計分野
  - 生物統計の本物の専門家を目指す人向け。最低1年間専念。

●いずれも昨年度実績。今年度について詳しくは：
http://www.niph.go.jp/soshiki/gijutsu/index_j.html
●このハンドアウトの最新版(6/9以降に更新)：
http://www.niph.go.jp/soshiki/jinzai/download/hotetsu2009/hotetsu2009.pdf