2019/5/9　北大SPH・統計解析の基礎⑦⑧

回帰分析

北海道大学　医学統計学
横田　勲

今回の内容

- 一般線形モデル
  - デザイン行列
  - 最小二乗法
  - 回帰診断
  - カテゴリカル変数のコーディング
- 一般線形混合効果モデル
- 一般化線形モデル
  - リンク関数と分布
- 関連、因果、予測
  - 決定係数、予測性能指標、ROC曲線

分散分析ではできないこと

- 因子は離散的なもののみ
  - 治療群、性別、・・・
- 年齢や血清マーカーのような
  連続量の影響も検討したい
  - 構造モデルを基にした
    一般線形モデル general linear model

一元配置の構造モデル

- $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$
  - y_{ij}：第i水準のj番目の対象者の反応
  - α_iを全体平均(μ)と第i水準の効果(α_i)に分解
    - $\sum_{i=1}^{a} \alpha_i = 0$という制約
  - ε_{ij}は互いに独立に
    平均0、分散σ^2の正規分布に従うと仮定
- 第i水準の平均値は$\mu + \alpha_i$として表現

回帰モデルで書き直し

- $y_{ij} = \mu + \alpha_1 x_1 + \cdots + \alpha_a x_a + \varepsilon_{ij}$
  - x_1：第1水準の場合に1, それ以外は0をとる
  - ⋮
  - x_a：第a水準の場合に1, それ以外は0をとる
- やっぱり第i水準の平均値は$\mu + \alpha_i$

各対象者の反応に対する回帰モデル

$$
\begin{aligned}
y_{11} &= \mu + \alpha_1 \times 1 + \alpha_2 \times 0 + \cdots + \alpha_a \times 0 + \varepsilon_{11} \\
&\vdots \\
y_{1n} &= \mu + \alpha_1 \times 1 + \alpha_2 \times 0 + \cdots + \alpha_a \times 0 + \varepsilon_{1n} \\
y_{21} &= \mu + \alpha_1 \times 0 + \alpha_2 \times 1 + \cdots + \alpha_a \times 0 + \varepsilon_{21} \\
&\vdots \\
y_{2n} &= \mu + \alpha_1 \times 0 + \alpha_2 \times 1 + \cdots + \alpha_a \times 0 + \varepsilon_{2n} \\
&\vdots \\
&\vdots \\
y_{a1} &= \mu + \alpha_1 \times 0 + \alpha_2 \times 0 + \cdots + \alpha_a \times 1 + \varepsilon_{a1} \\
&\vdots \\
y_{an} &= \mu + \alpha_1 \times 0 + \alpha_2 \times 0 + \cdots + \alpha_a \times 1 + \varepsilon_{an}
\end{aligned}
$$

行列表現

7

$$\begin{pmatrix} y_{11} \\ \vdots \\ y_{1n} \\ y_{21} \\ \vdots \\ y_{2n} \\ \vdots \\ y_{a1} \\ \vdots \\ y_{an} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 1 & 0 & & 0 \\ 1 & 0 & 1 & & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 1 & & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & & 1 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_a \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1n} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{2n} \\ \vdots \\ \varepsilon_{a1} \\ \vdots \\ \varepsilon_{an} \end{pmatrix}$$

$\boldsymbol{Y}$ $\boldsymbol{X}$ $\boldsymbol{\beta}$ $\boldsymbol{\varepsilon}$

次元 $an \times 1$ $an \times (a+1)$ $(a+1) \times 1$ $an \times 1$

一般線形モデル

8

- $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$
  - アウトカム $\boldsymbol{y}$ を
    モデルで説明する分 $\boldsymbol{X\beta}$ と誤差 $\boldsymbol{\varepsilon}$ の和で表現
  - $\boldsymbol{X}$：デザイン行列 design matrix
  - $\boldsymbol{\beta}$：回帰係数(ベクトル) coefficient (vector)

デザイン行列

9

- 作り方次第で、以下の解析を表現可能
  - 線形回帰分析（単回帰、重回帰）
  - t検定
  - 一元配置分散分析
  - 多元配置分散分析（ブロック因子を含む）
  - 繰り返し測定分散分析
  - 共分散分析 analysis of covariance
  - ・・・

例：t検定を表現①

10

- 群1と群2の平均値を比較したい
- $y_{ij} = \beta_0 + x_1\beta_1 + \varepsilon_{ij}$
  - x_1：群1なら1,群2なら0
- 群1の平均値

$$\frac{y_{11} + y_{12} + \cdots + y_{1n}}{n}$$
$$= \frac{(\beta_0 + 1 \times \beta_1 + \varepsilon_{11}) + (\beta_0 + 1 \times \beta_1 + \varepsilon_{12}) + \cdots + (\beta_0 + 1 \times \beta_1 + \varepsilon_{1n})}{n}$$
$$= \beta_0 + \beta_1$$

（∵ランダム誤差は期待値が0）

例：t検定を表現②

11

- 群2の平均値

$$\frac{y_{21} + y_{22} + \cdots + y_{2n}}{n}$$
$$= \frac{(\beta_0 + 0 \times \beta_1 + \varepsilon_{21}) + (\beta_0 + 0 \times \beta_1 + \varepsilon_{22}) + \cdots + (\beta_0 + 0 \times \beta_1 + \varepsilon_{2n})}{n}$$
$$= \beta_0$$

- 群間差(群1の平均値ー群2の平均値)はβ_1
- $\beta_1 = 0$であるかの検定がt検定

記法

12

- 分散分析とは記法をやや変更
- $\boldsymbol{Y} = (y_1, y_2, \cdots, y_n)^\top$
  - y_i：対象者 $i(= 1, \cdots, n)$ のアウトカム
- $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots)^\top$
  - β_0：切片項 intercept の回帰係数

例：単回帰分析

13

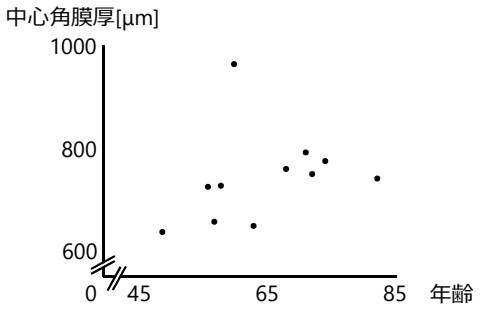
- 水疱性角膜症の術前中心角膜厚

ID	年齢[歳]	術前中心角膜厚[μm]
1	68	760
2	60	964
3	58	727
4	71	792
5	49	637
6	74	775
7	72	750
8	57	657
9	63	649
10	82	741
11	56	725

Kinoshita S, *et al. N Engl J Med*. 2018; 995-1003.

散布図

14



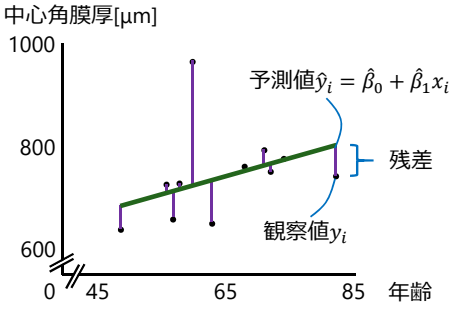
角膜厚と年齢の関連

15

- 直線的な関係があるか？
  - 年齢が1歳上がるごとに、角膜厚は平均的にどれだけ変化するか
- 角膜厚と年齢の関係を表す最も適切な直線を求めたい
  - $y_i = \beta_0 + x_i\beta_1 + \varepsilon_i$
    - β_0：切片
    - β_1：傾き

残差平方和を最小に

16



最小二乗法

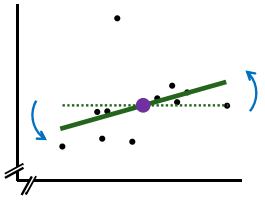
17

- $\sum_i^n (y_i - \hat{y}_i)^2 = \sum_i^n \{y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)\}^2$ を最小にする$\hat{\beta}_0, \hat{\beta}_1$
  - 推定量や推定値について ^（ハット）を付す
  - $\begin{cases} \frac{\partial \sum_i^n (y_i - \hat{y}_i)^2}{\partial \hat{\beta}_0} = 0 \\ \frac{\partial \sum_i^n (y_i - \hat{y}_i)^2}{\partial \hat{\beta}_1} = 0 \end{cases}$ を解く
- 一般的には$\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^\top \boldsymbol{Y}$を解く
  - 正規方程式

単回帰の場合

18

- ① yとxの平均をみつける
- ② 平均値を中心に直線を回転し、残差平方和最小となる場合をみつける



誤差の4条件 19

- 独立性
  - 各対象者の誤差が互いに独立
- 不偏性
  - 誤差の期待値は0
- 等分散性
  - 誤差分散は等しい
- 正規性
  - 誤差は正規分布にしたがう
- $\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$

推定量の特徴 20

- 誤差について、
  正規性を除く3条件を仮定した下で、
  最小二乗法による回帰係数の推定量は
  - 不偏
    - 十分にnが大きい下で、バイアスがない
  - 最小分散（有効推定量）

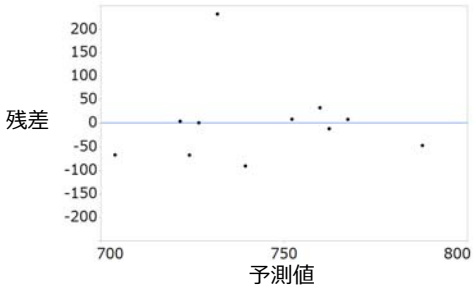
  を満たす（最良線形不偏推定量;BLUE）

回帰診断 21

- 一般線形モデルの前提が正しいか
  - 説明変数の関数形
    - 効果の加法性、交互作用の検討
  - 誤差の4(3)条件
- 残差分析
- 影響度分析

残差分析 22

- 残差を予測値や説明変数に対してプロット



影響度分析 23

- ある人のデータを除いたときに、
  回帰係数の推定値や予測値が
  どれくらい変化するか
- Cookの距離
  - 予測値の変化を基準化した指標

デザイン行列の作り方 24

- 3水準、繰り返し数が2回の一元配置

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{pmatrix}$$

$\boldsymbol{Y}$ $\boldsymbol{X}$ $\boldsymbol{\beta}$ $\boldsymbol{\varepsilon}$

- 正規方程式を解いてみよう

$X^\top X\beta = X^\top Y$ 25

- $X^\top X$

$$= \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 6 & 2 & 2 & 2 \\ 2 & 2 & 0 & 0 \\ 2 & 0 & 2 & 0 \\ 2 & 0 & 0 & 2 \end{pmatrix}$$

$X^\top X\beta = X^\top Y$ 26

- $$\begin{cases} 6\beta_0 + 2\beta_1 + 2\beta_2 + 2\beta_3 = y_1 + \cdots + y_6 \\ 2\beta_0 + 2\beta_1 = y_1 + y_2 \\ 2\beta_0 + 2\beta_2 = y_3 + y_4 \\ 2\beta_0 + 2\beta_3 = y_5 + y_6 \end{cases}$$
- 解が不定
  - 3水準の実験から
    4つの平均値を求めるなんてムリ

一意に推定可能な次数に 27

- 切片をなくす
  - GLM coding without intercept
- ある水準を参照水準とする
  - Reference coding
- ある水準との対比
  - Effect coding

各コーディング 28

GLM coding

	x_1	x_2	x_3
水準1	1	0	0
水準2	0	1	0
水準3	0	0	1

Reference coding

	x_1	x_2
水準1	1	0
水準2	0	1
水準3	0	0

Effect coding

	x_1	x_2
水準1	1	0
水準2	0	1
水準3	-1	-1

GLM coding without intercept 29

- $E(y) = x_1\beta_1 + x_2\beta_2 + x_3\beta_3$
  - 水準1の平均値
    $E(y) = 1 \cdot \beta_1 + 0 \cdot \beta_2 + 0 \cdot \beta_3 = \beta_1$
  - 水準2の平均値
    $E(y) = 0 \cdot \beta_1 + 1 \cdot \beta_2 + 0 \cdot \beta_3 = \beta_2$
  - 水準3の平均値
    $E(y) = 0 \cdot \beta_1 + 0 \cdot \beta_2 + 1 \cdot \beta_3 = \beta_3$

Reference coding 30

- $E(y) = \beta_0 + x_1\beta_1 + x_2\beta_2$
  - 水準1の平均値
    $E(y) = 1 \cdot \beta_0 + 1 \cdot \beta_1 + 0 \cdot \beta_2 = \beta_0 + \beta_1$
  - 水準2の平均値
    $E(y) = 1 \cdot \beta_0 + 0 \cdot \beta_1 + 1 \cdot \beta_2 = \beta_0 + \beta_2$
  - 水準3（参照水準）の平均値
    $E(y) = 1 \cdot \beta_0 + 0 \cdot \beta_1 + 0 \cdot \beta_2 = \beta_0$
- β_0：水準3の平均値
- β_1：水準1の平均値 – 水準3の平均値
- β_2：水準2の平均値 – 水準3の平均値

Effect coding

31

- $E(y) = \beta_0 + x_1\beta_1 + x_2\beta_2$
  - 水準1の平均値
    $E(y) = 1 \cdot \beta_0 + 1 \cdot \beta_1 + 0 \cdot \beta_2 = \beta_0 + \beta_1$
  - 水準2の平均値
    $E(y) = 1 \cdot \beta_0 + 0 \cdot \beta_1 + 1 \cdot \beta_2 = \beta_0 + \beta_2$
  - 水準3（参照水準）の平均値
    $E(y) = 1 \cdot \beta_0 - 1 \cdot \beta_1 - 1 \cdot \beta_2 = \beta_0 - \beta_1 - \beta_2$
- β_0：全体平均
  - (水準1の平均値 + 水準2の平均値 + 水準3の平均値)/3
- β_1：水準1の平均値 – 全体平均
- β_2：水準2の平均値 – 全体平均

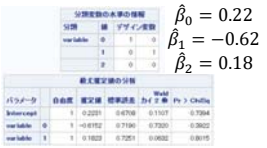
SASでの例

32

GLM coding without intercept

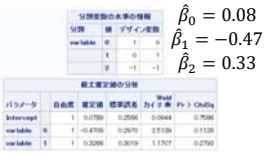
$\hat{\beta}_1 = -0.39$
$\hat{\beta}_2 = 0.41$
$\hat{\beta}_3 = 0.22$

Reference coding



$\hat{\beta}_0 = 0.22$
$\hat{\beta}_1 = -0.62$
$\hat{\beta}_2 = 0.18$

Effect coding



$\hat{\beta}_0 = 0.08$
$\hat{\beta}_1 = -0.47$
$\hat{\beta}_2 = 0.33$

説明変数を自分でコーディング

33

- 解釈しやすい回帰係数となるように
- 例：JMPでロジスティック回帰
  - 説明変数(trt)を0,1で入力

説明変数を名義変数

パラメータ推定値

項	推定値	標準誤差
切片	-3.7290143	0.1563322
trt[0]	0.06545268	0.1563322

推定値は次の対数オッズに対するもの

Effectコーディング

説明変数を連続変数

パラメータ推定値

項	推定値	標準誤差
切片	-3.6635616	0.2264554
trt	-0.1309054	0.3126645

推定値は次の対数オッズに対するも

相関のあるデータ

34

- アウトカムが互いに独立という前提が成立しない
  - 同一対象者、マッチしたペア、施設等
- 相関が生じる単位を「ブロック」とした
  - ブロック因子を含む二元配置分散分析
  - 繰り返し測定分散分析
  - Unbalancedな2因子実験

例：中心角膜厚データ

35

- 水準数3、ブロック数11の乱塊法

ID	術前[μm]	術後24週	術後2年
1	760	511	532
2	964	525	538
3	727	540	546
4	792	640	710
5	637	509	529
6	775	505	525
7	750	626	550
8	657	489	503
9	649	595	543
10	741	539	523
11	725	561	572

Kinoshita S, *et al. N Engl J Med*. 2018; 995-1003.

線形回帰モデルで表現

36

- $y_i = \beta_0 + \underbrace{x_1\beta_1 + x_2\beta_2 + x_3\beta_3}_{\text{時点効果}} + \underbrace{z_1\gamma_1 + z_2\gamma_2 + \cdots + z_{11}\gamma_{11}}_{\text{ブロック因子}} + \varepsilon_i$
  - x_1：時点が術前なら1、それ以外は0
  - x_2：時点が24週なら1、それ以外は0
  - x_3：時点が2年なら1、それ以外は0
  - z_1：IDが1なら1、それ以外は0
  - ⋮
  - z_{11}：IDが11なら1、それ以外は0

ブロック因子もすべて推定する必要がある

繰り返し測定分散分析の構造モデル 37

- $y_{ijk} = \mu + \alpha_i + u_{ij} + \beta_k + (\alpha\beta)_{ik} + \varepsilon_{ijk}$
  - μ ：全体平均
  - α_i ：治療の主効果（①）
  - u_{ij} ：1次誤差
    - 治療の割付をした個人における誤差
  - β_k ：時点の主効果（②）
  - $(\alpha\beta)_{ik}$ ：治療と時点の交互作用（③）
  - ε_{ijk} ：2次誤差
    - 個人のうち各時点の測定に関する誤差

ブロックの代わりに誤差を入れた

2つの誤差 38

- 個人による誤差
  - ヒトの違いによって生じる
  - 同一対象者では同じ値を持つが、対象者間では値が分布する
  - 変量効果 random effect と呼ぼう
- 個人内での繰り返し測定による誤差
  - 同じヒトでも複数回測定したことによって生じる

固定効果と変量効果 39

- 固定効果 fixed effects
  - 因子の水準に普遍性があり、効果自体の推定に関心がある
    - 治療、時期、予後因子・・・
- 変量効果 random effects
  - 因子の水準はある集団からのサンプルと考え、効果自体の推定ではなく、そのバラツキの大きさに関心がある
    - 測定単位（個人、マッチしたペア、施設など）に特有の効果

混合効果モデルの別名 40

- 混合モデル mixed model
  - mixture modelは全く別の意味！
- 変量効果モデル random effect model
- マルチレベルモデル multi-level model
  - 社会科学の分野で一般的

ランダム係数モデル 41

- $y_{ij} = \alpha + Time \times \beta + b_{0i} + Time \times b_{1i} + \varepsilon_{ij}$
  - y_{ij}：対象者 i の時点 j でのアウトカム
  - 変量切片b_{0i}に加え、時点との交互作用変量b_{1i}を加え、経時変化の傾きについても個人差をゆるす
    - b_{0i}とb_{1i}は二変量正規分布に従う

混合効果モデルの特徴 42

- 個人差を適当に調整して、個人全体でのアウトカム評価を行いたい
  - 繰り返し測定分散分析に比べ、検出力を向上できる場面がある
- 対象者間での測定回数の違いにもある程度対応可能
  - 1回しか測定されない人と複数回測定される人が含まれる集団でもOK

モデルパラメータ推定

43

- 推奨
  - 制限付き最尤(REML)法
  - 分散はサンドウィッチ推定量
    - JMP14.0では対応していない・・・
- 要検討
  - 変量効果の相関構造
  - 欠測に対する仮定
  - 自由度の調整
  - ・・・

効果の指標

44

- 治療(曝露)効果の方向や大きさの表現

指標	差の指標	比の指標
リスク、割合	リスク差	リスク比
オッズ		オッズ比
率	率差	率比
ハザード		ハザード比

2つの治療法を比較する1800名の観察研究

45

治療法	イベントあり	イベントなし	合計
試験治療	22 (2.2%)	978	1000
標準治療	20 (2.5%)	780	800
合計	42	1758	1800

- 標準治療に対する試験治療の
  - リスク差：$2.2\% - 2.5\% = -0.3\%$
  - リスク比：$2.2\%/2.5\% = 0.88$
  - オッズ比：$\frac{22/978}{20/780} = 0.877\ldots \approx 0.88$

リスク差の回帰モデル①

46

- 一般線形モデルで行ったら？
  - t検定のように回帰係数を平均値の差となるようモデルを作れたし・・
- アウトカムは連続量ではなく、イベントあり／なしの二値
  - アウトカムは二項分布に従うと考えたほうが自然

リスク差の回帰モデル②

47

- $y_i \sim Bin(n_i, p_i)$
  - アウトカムは二項分布に従う
  - 今は$n_i = 1$
  - イベント発生確率がp_i
  - $E(y_i) = n_i p_i$
    - 今はy_iの期待値がイベント発生確率p_i
- $p_i = \beta_0 + x_i\beta_1$
  - x_i：試験治療なら1、標準治療なら0

リスク差の回帰モデル③

48

パラメータ	推定値	(95%信頼区間)
β_0	0.025	(0.014, 0.036)
β_1	-0.003	(-0.017, 0.011)

- 試験治療により、イベント発生が-0.3%(95%CI: -1.7%, 1.1%)だけ減る
  - NNTは$\frac{1}{|-0.003|} \approx 333$
- 割合の差の検定に基づく信頼区間と同じ

一般化線形モデル generalized linear model 49

- アウトカムが指数型分布族に従う
  - 正規分布、二項分布、Poisson分布など
- 分布を表現する正準パラメータθについて
$$g(\theta) = \boldsymbol{X\beta}$$
- $g(\cdot)$　リンク関数
  - リスク差モデルではそのまま$g(\theta) = \theta$

分布とリンク関数の組合せ 50

分布	リンク関数	効果の指標	別名
二項	恒等	リスク差	
二項	対数	リスク比	
二項 / 多項	ロジット	オッズ比	ロジスティック回帰
Poisson	恒等	率差	
Poisson	対数	率比	Poisson回帰
指数	対数	ハザード比	指数回帰
二項	プロビット		プロビット回帰

- 医学研究ではCox回帰を用いたハザード比推定がほとんど
- イベント割合が極端に低い場合に向かない、効果指標がないといった理由でプロビット回帰はあまり使われない

指数関数・対数関数 51

- $\log(A \times B) = \log A + \log B$
  - 対数の底はeであり、通常省略
  - 比のモデルは掛け算モデル
対数変換すれば足し算モデルになる
    - 線形モデルで表現できる
- $e^{\log A} = A$
- $e^{a+b} = e^a \times e^b$
  - 以降、e^aのことを$\exp a$と表記する
    - $\exp(a+b) = \exp a \times \exp b$

1800名のデータ 52

- 年齢でサブグループ化
- 75歳以上

治療法	イベントあり	イベントなし	合計
試験治療	18 (3.0%)	582	600
標準治療	6 (6.0%)	94	100

- 75歳未満

治療法	イベントあり	イベントなし	合計
試験治療	4 (1.0%)	396	400
標準治療	14 (2.0%)	686	700

ロジスティック回帰の例 53

オッズ

- $\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2$
  - x_{1i}：試験治療なら1、標準治療なら0
  - x_{2i}：75歳以上なら1、75歳未満なら0
- 交互作用項は含めていないので、
年齢によらず治療効果は同じという仮定

ロジスティック回帰分析結果 54

	β	(95%CI)	オッズ比	(95%CI)
切片	-3.89	(-4.38, -3.40)		
x_1；治療	-0.72	(-1.44, 0.01)	0.49	(0.28-1.01)
x_2；年齢	1.13	(0.40, 1.86)	3.10	(1.50-6.49)

- 治療のオッズ比とその信頼区間
$$e^{-0.72} = \exp(-0.72) = 0.49$$
$$e^{-1.44} = \exp(-1.44) = 0.28$$
$$e^{0.01} = \exp(0.01) = 1.01$$

年齢を調整した治療のオッズ比 55

$\log(\text{オッズ}) = \beta_0 + \beta_1 \times (\text{試験治療}) + \beta_2 \times (75\text{歳以上})$

オッズ	75歳未満	75歳以上
試験治療	$\exp(\beta_0 + \beta_1)$	$\exp(\beta_0 + \beta_1 + \beta_2)$
標準治療	$\exp(\beta_0)$	$\exp(\beta_0 + \beta_2)$

75歳未満でのオッズ比
$$\frac{\exp(\beta_0 + \beta_1)}{\exp(\beta_0)} = \exp(\boldsymbol{\beta_1})$$

75歳以上でのオッズ比
$$\frac{\exp(\beta_0 + \beta_1 + \beta_2)}{\exp(\beta_0 + \beta_2)} = \exp(\boldsymbol{\beta_1})$$

練習問題 56

- 75歳未満で試験治療を受けた人に対する75歳以上で標準治療を受けた人のイベント発生オッズ比は？

医学研究での目的 57

- "X" は "疾病Y" と 関連 がある
  - X： 健康状態マーカーや疾病Yを引き起こす疾患など

↓

- "X" は "疾病Y" の 原因 となる
- "X" は "疾病Y" を 予測 する

より目的を明確に

因果と予測 58

- 回帰分析から、X-Y間の「関連」を検討
- Xが原因となり、Yという結果が導かれる
  - 回帰モデルは因果モデル (`do' model)
  - 交絡因子は制御すべきもの
- Xの値を与えて、Yという結果を当てる
  - 回帰モデルは予測モデル (`see' model)
  - 予測精度を高めるためにXを選ぶ

Allison PD. 1998(Book).
vanHouwelingen JC. The President's speech in ISCB34.

前立腺がんとPSA 60

- 前立腺がんの発見・病勢と強い関連
  - スクリーニングにも用いられる
- がんの細胞壁が壊れやすいため、がんのvolumeに応じてPSAが血液中に漏出

前立腺がん → PSA

前立腺がんのリスク因子を検討 61

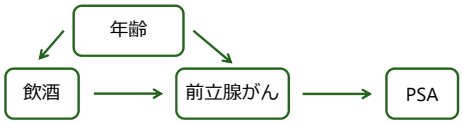
- 明らかなリスク因子は、年齢、家族歴
- 他にもリスク因子はあるに違いない！
  - 例えば、飲酒の影響を調べてみる
  - 因果関係を知りたい

飲酒 → 前立腺がん → PSA

交絡の影響を解析で除去

62

- 以下の条件を満たすことで、
  飲酒と前立腺がんの関係を歪めてしまう
  - 年齢が高いほど前立腺がんは増える
  - 年齢と飲酒には関係がある
  - 飲酒をすれば年齢が増えるわけではない



たいてい多変量解析で除去

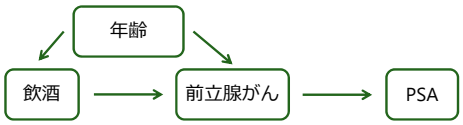
63

Sawada N. et al. *Int J Cancer*. 2014.

前立腺がんの予測をしたい

64

- 前立腺がん発生を精度良く当てたい
  - どのような因子を用いてもよい
    - 年齢のようなリスク因子
    - 前立腺がんの"結果"であるPSA



http://psacalc.sph.umich.edu

65

変数選択

66

- 予測モデルを作るため
  - 少ない変数で当たりのよいモデルを
    - 特別な測定を要する変数で作ったモデルは
      使われづらい
- 因果関係を調べるためには使わない
  - 交絡調整が目的ゆえ、
    利用可能なすべての変数を用いる

モデルのよさ、精度の測り方

67

- モデルのあてはまり
  - 決定係数R^2
  - 尤度とAIC
- 予測精度、予測結果のよさ
  - 平均二乗誤差、Brierスコア
  - ROC曲線、c 統計量

決定係数 R^2 68

残差平方和　モデルで説明した平方和

- $R^2 = 1 - \frac{(y_i - \hat{y}_i)^2}{(y_i - \bar{y})^2} = \frac{(\hat{y}_i - \bar{y})^2}{(y_i - \bar{y})^2}$

全体の平方和

- データの持つ全ばらつきのうち、モデルで説明した割合
  - 単回帰の場合、相関係数の2乗に一致

確率(密度)関数と尤度 69

- データは確率変数の実現値
  - 確率分布（パラメータ$\boldsymbol{\beta}$を持つモデル）を仮定すれば、当該データが得られる確からしさを定義
  - $f(x_1, \dots, x_n; \boldsymbol{\beta}) = \prod_i^n f(x_i; \boldsymbol{\beta})$
- 尤度 $L(\boldsymbol{\beta}; x)$
  - $f(x; \boldsymbol{\beta})$を$\boldsymbol{\beta}$の関数としてみたもの

最尤法 maximum likelihood method 70

- 尤度が最大になるようなパラメータθを推定値とみなす手法
  - 当該データが得られる確からしさが最大ゆえ
- 例：一般線形モデル
  - 誤差に正規分布を仮定すれば、最尤推定値と最小二乗法による推定値が一致
- 例：一般化線形モデル
  - たいていの場合、最尤推定量は統計的によい性質をもつ

尤度によるあてはまりの評価 71

- 連続量アウトカム以外では尤度で考える
  - 大きいほどデータへのあてはまりがよい
  - 誤差に正規分布を仮定した場合、幸いにも、誤差平方和が小さくなるほど尤度は大きくなる関係

モデルの複雑さとoverfitting 72

- モデルを複雑にすると より細かな違いまで捉えられる
  - 関数形を高次にする
  - 説明変数を増やす
    - それが無意味な説明変数であっても！
- 無意味な説明変数をモデル化すると、他のデータへのあてはまりが悪くなる
  - モデルを活用する際に使いづらい
  - Overfitting （過適合）という

Akaike's Information Criterion; AIC 73

- $-2 \log L(\boldsymbol{\beta}; x) + 2K$
  - K ：$\boldsymbol{\beta}$ のパラメータ数
- AICが最小となるモデルがよいモデル
  - パラメータを増やすことへのペナルティを与えた指標
    - 自由度調整済み決定係数も同様

古典的な変数選択法 74

- 基準に至るまで以下の操作を繰り返す
- 変数増加法
  - 変数候補から最もp値の小さなものを加える
- 変数減少法
  - 変数候補をすべて含めたモデルから最もp値の大きな変数を除く
- ステップワイズ法
  - 変数候補から最もp値の小さなものを加え、モデルから最もp値の大きな変数を除く

他の変数選択法 75

- p値の代わりに用いる基準
  - AIC
  - 平均二乗誤差、Brierスコア
  - c-index
  - ・・・
- 総当たり法
  - 変数の組合せ全パターン調べる

平均二乗誤差 Mean Squared Error 76

- $\frac{1}{n}(\hat{y}_i - y_i)^2$
  - 予測値と実測値の差を評価
  - 平方根をとって、Root MSE; RMSE

Brier スコア 77

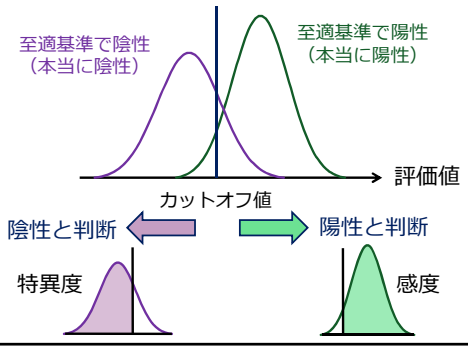
- イベント有無と生存確率のズレ
  - 生存時間アウトカムの場合、ある時点tでのイベント有無と確率のズレ
- Brierスコア
  - $\{I(y = 1) - \hat{y}\}^2$
  - $\left\{I(T > t) - \hat{S}(t|\boldsymbol{X})\right\}^2$
    - $I(\cdot)$：かっこ内が真のときに1、それ以外は0
    - $\hat{S}(\cdot)$：生存関数の予測値

感度と特異度 78

	評価値	
至適基準	陽性	陰性
陽性	a	b
陰性	c	d

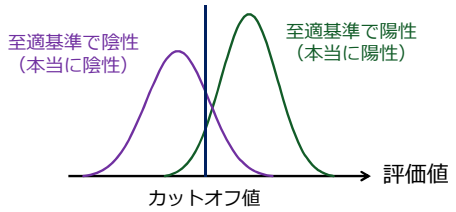
- 感度：$\frac{a}{a+b}$
  - 本当に陽性であるものを陽性といえたか
- 特異度：$\frac{d}{c+d}$
  - 本当に陰性であるものを陰性といえたか

カットオフ値をもって評価 79



感度と特異度はトレードオフ

80

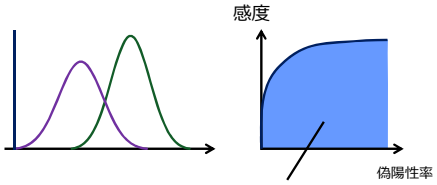


- 感度を上げれば、特異度は下がる
- 特異度を上げれば、感度は下がる

ROC曲線

81

- 縦軸に感度、横軸に偽陽性率（1-特異度）
- カットオフ値を全範囲で動かした場合の感度と偽陽性率をプロット



AUC (Area Under the Curve)

(ROC-)AUC

82

- ROC曲線の要約指標
  - 判別能力を表す指標として解釈
- AUC自体はモデルに依らずに計算される
  - AUC=0.5であれば、no discriminative ability
  - AUCが1に近づくほど、判別能力がよい
  - 絶対値的な解釈は困難
- c (concordance) indexとも呼ばれる

c-index

83

$$\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\{I(\hat{p}_i > \hat{p}_j) + 0.5 \times I(\hat{p}_i = \hat{p}_j)\}}{nm}$$

- $\hat{p}_i$：実際にイベント発生したiの イベント確率の推定値
- $\hat{p}_j$：実際にイベント発生しなかったjの イベント確率の推定値
- イベント有無と予測変数の
  - 大小関係が一致していれば1点
  - 大小関係が一致していなければ0点
  - 値が等しければ0.5点（引き分け）

陽性的中度、陰性的中度との違い

84

	評価値	
至適基準	陽性	陰性
陽性	a	b
陰性	c	d

- 陽性的中度：$\frac{a}{a+c}$、陰性的中度：$\frac{d}{b+d}$
  - 評価した結果が本当はどうであったか？
  - 真の陽性、陰性者の分布によって 変わってしまう指標

c-indexを生存時間データに拡張

85

- 生存時間そのもの；overall C
  - 生存時間が短い（予後が悪い）人ほど、予測変数が大きな値であればconcordant
    - 検討する時間の範囲を設けてもよい；dynamic C
- Uno's cが標準的

Uno H, *et al. Stat Med.* 2011. 1105-1117.

まとめ①

86

- 一般線形モデル
  - デザイン行列をうまく作って、
    t検定、分散分析、回帰分析等を統一的に表現
  - 残差平方和を最小にする最小二乗法
  - カテゴリカル変数のコーディング
    - GLM / reference / effect coding
    - コーディングによって推定値の解釈に注意

まとめ②

87

- 一般線形混合モデル
  - 変量効果によって個別の平均に興味はないが、
    ばらつきを生むものを表現
- 一般化線形モデル
  - リスク差・比回帰、ロジスティック回帰、
    Poisson回帰等を統一的に表現
  - アウトカムの従う分布を定め、
    分布のパラメータを線形モデルで表現
    - リンク関数によって柔軟にモデル化

まとめ③

88

- 因果モデルと予測モデル
  - 変数選択の考え方に違い
  - 決定係数、AIC
    - 一般化線形モデルの推定は最尤法
  - 平均二乗誤差、Brierスコア
  - ROC-AUC、c-index