2021/1/21　医療統計学（３年）⑤

回帰分析

北海道大学　医学統計学
横田　勲

1

今回の内容

2

- 一般線形モデル
- ロジスティック回帰モデル

2

例：ラットの反復投与実験

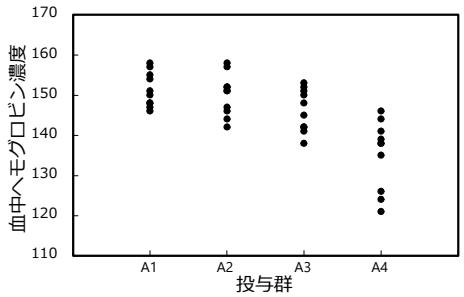
3

- ある薬物Aの影響を調べるために、ラット40匹を10匹ずつの4群に分け、薬剤の静脈内投与を35日続けた
  - A1 : 対照群　　A2 : 薬物Aを5mg/kg
  - A3 :薬物Aを10mg/kg　A4 : 薬物Aを20mg/kg
- 反応：血中ヘモグロビン濃度 (mg/dl)
- 因子：薬物濃度（投与群）
- 水準：A1, A2, A3, A4の4水準

3

ヘモグロビン濃度の分布

4



4

線形回帰分析

5

- 対象者$k(=1,\dots,n)$に対する単回帰分析
  - i番目の対象者の反応 Y_k
  - 説明変数 X_k
    - 治療群、性別、年齢、血清マーカー、・・・
  - 回帰パラメータ α,β
  - ランダム誤差 $\varepsilon_k \sim N(0,\sigma^2)$
- 以下の足し算を用いた方程式を置く

$$Y_k = \alpha + X_k\beta + \varepsilon_k$$

5

一般線形モデル general linear model

6

- 回帰分析のみならず、t検定や分散分析も統一的に扱うことが可能
- t検定の場合
  - 治療群が1のとき$X_k = 1$、群が2のとき$X_k = 0$

$$Y_k = \alpha + X_k\beta + \varepsilon_k$$

  - 誤差ε_kは期待値がゼロ

6

期待値をとってみよう 7

- 群1であるkさん　$Y_k = \alpha + 1 \times \beta + \varepsilon_k$
- 群1全員の期待値（平均値）は、

$$\frac{(\alpha + \beta + \varepsilon_1) + (\alpha + \beta + \varepsilon_2) + \cdots}{n_1} = \alpha + \beta$$

（∵ランダム誤差は期待値が0）

- 同様に、群2の期待値は、

$$\frac{(\alpha + \varepsilon_1) + (\alpha + \varepsilon_2) + \cdots}{n_2} = \alpha$$

7

t検定の場合 8

- $Y_k = \alpha + X\beta + \varepsilon_k$
  - α：群2の平均値
  - β：群間差
- $\beta = 0$であるかの検定がt検定

8

一元配置分散分析の場合 9

- 治療群が4つ(1,2,3,4)あるならば、
  - 治療群が1のとき$X_1 = 1$、それ以外$X_1 = 0$
  - 治療群が2のとき$X_2 = 1$、それ以外$X_2 = 0$
  - 治療群が3のとき$X_3 = 1$、それ以外$X_3 = 0$
  - 治療群が4のとき$X_4 = 1$、それ以外$X_4 = 0$

$$Y_k = X_1\beta_1 + X_2\beta_2 + X_3\beta_3 + X_4\beta_4 + \varepsilon_k$$

- $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$であるかの検定が$F$検定

9

バラツキ variation 10

- 同じようなものを測定したはずなのに、値が異なってしまうこと
  - 統計で問題にするのはこの「バラツキ」
- 頻度流統計学では

測定値 (measurement) ＝ 真値 (true value) ＋ バイアス (bias) ＋ 誤差 (error)

腎摘出術を受ける患者の真のeGFR

年齢・性別による影響
前治療による影響
⋮

10

バラツキの分解という観点 11

- 一般線形モデルでも同じ
- でも真値は分からない
  - そもそもあまり興味もない
  - とりあえず基準点を置いて、切片と呼ぼう
    - 全てモデル因子で分解して、切片を置かなくたっていい

測定値 (measurement) ＝ 切片 (intercept) ＋ モデル因子 (model factor) ＋ 誤差 (error)

$$Y_k = \alpha + X_k\beta + \varepsilon_k$$

11

モデルの構築 12

- 例えば、対照群との比較に興味があれば

$$Y_k = \alpha + X_1\beta_1 + X_2\beta_2 + X_3\beta_3 + \varepsilon_k$$

  - X_1：治療群がA2の時だけ1、それ以外は0
  - X_2：治療群がA3の時だけ1、それ以外は0
  - X_3：治療群がA4の時だけ1、それ以外は0

12

解析結果の見方① 13

- 解析ソフトSASの出力

パラメータ	推定値	標準誤差	t 値	Pr > |t|	95% 信頼限界	
Intercept	151.4000000	1.92757764	78.54	<.0001	147.4906914	155.3093086
A2	−1.4000000	2.72600644	−0.51	0.6107	−6.9285973	4.1285973
A3	−5.2000000	2.72600644	−1.91	0.0645	−10.7285973	0.3285973
A4	−16.2000000	2.72600644	−5.94	<.0001	−21.7285973	−10.6714027

- 整形して

治療群	推定値	95%信頼区間	両側P値
A2 vs A1	-1.4	(-6.9, 4.1)	0.61
A3 vs A1	-5.2	(-10.7, 0.3)	0.06
A4 vs A1	-16.2	(-21.7, -10.7)	<0.01

13

解析結果の見方② 14

治療群	推定値	95%信頼区間	両側P値
A2 vs A1	-1.4	(-6.9, 4.1)	0.61
A3 vs A1	-5.2	(-10.7, 0.3)	0.06
A4 vs A1	-16.2	(-21.7, -10.7)	<0.01

- 治療群A2はA1に比べ、
  ヘモグロビン濃度の平均の差は-1.4
  - A2のヘモグロビン濃度のほうが低い
  - その信頼区間は(-6.9, 4.1)

14

薬物濃度を連続量として扱う 15

- 薬物Aを与えるごとに、
  どれだけヘモグロビン濃度は低下するか？
  - A1～A4は0,5,10,20mg/kg投与
  - 薬物増加に従い、
    ヘモグロビン濃度は直線的に変化すると仮定
- Xは薬物Aの濃度として、モデルを構築

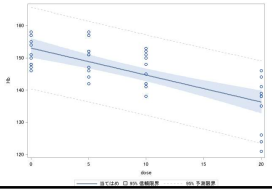
$$Y_k = \alpha + X\beta + \varepsilon_k$$

15

回帰直線の推定 16

パラメータ	推定値	標準誤差	t 値	Pr > |t|	95% 信頼限界	
Intercept	153.0400000	1.49372220	102.46	<.0001	150.0161175	156.0638825
dose	−0.8388571	0.13038276	−6.43	<.0001	−1.1028032	−0.5749110

- 薬物を1mg/kg増やすごとに、
  ヘモグロビン濃度は-0.84mg/dLだけ変化



16

線形回帰モデル 17

- 反応Yの変化を、
  説明変数（共変量）X で説明
  - 説明変数が群（カテゴリカル）であれば、
    群間差が求まる
  - 説明変数が連続量であれば、
    1単位変化あたりの反応の変化が求まる

17

例：単回帰分析 18

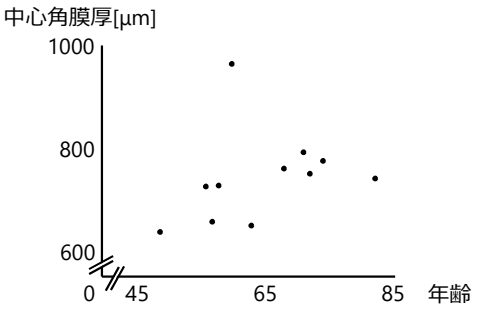
- 水疱性角膜症の術前中心角膜厚

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.

18

散布図 19



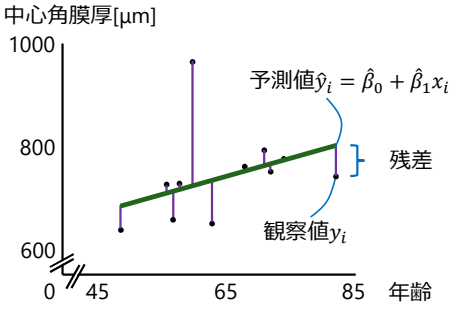
19

角膜厚と年齢の関連 20

- 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

20

残差平方和を最小に 21



21

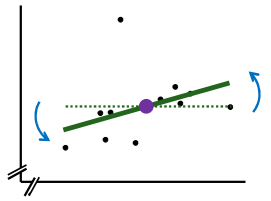
最小二乗法 22

- $\sum_i^n (y_i - \hat{y}_i)^2 = \sum_i^n \{y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)\}^2$ 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

22

単回帰の場合 23

- 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



23

連続量の回帰モデル 24

- ここまで扱った反応Yは連続量のみ
  - 反応あり／なしのような二値データや、人時間データでも行いたい

24

効果の指標

25

- 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

25

2つの治療法を比較する1800名の観察研究

26

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：$\frac{22/978}{20/780} = 0.877\ldots \approx 0.88$

26

リスク差の回帰モデル①

27

- 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

27

リスク差の回帰モデル②

28

- $y_i \sim Bin(n_i, p_i)$
  - 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

28

リスク差の回帰モデル③

29

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$\frac{1}{\lvert-0.003\rvert} \approx 333$
- 割合の差の検定に基づく信頼区間と同じ

29

一般化線形モデル generalized linear model

30

- 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$g(\theta) = \theta$

30

分布とリンク関数の組合せ 31

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

31

指数関数・対数関数 32

- $\log(A \times B) = \log A + \log B$
  - 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
    - $\exp(a+b) = \exp a \times \exp b$

32

1800名のデータ 33

- 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

33

ロジスティック回帰の例 34

- $\mathrm{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2$（$\frac{p_i}{1-p_i}$：オッズ）
  - 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

34

ロジスティック回帰分析結果 35

	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

$$e^{-0.72} = \exp(-0.72) = 0.49$$
$$e^{-1.44} = \exp(-1.44) = 0.28$$
$$e^{0.01} = \exp(0.01) = 1.01$$

35

年齢を調整した治療のオッズ比 36

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

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 $\frac{\exp(\beta_0 + \beta_1)}{\exp(\beta_0)} = \exp(\beta_1)$

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

36

練習問題 37

- 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？

37

まとめ 38

- 一般線形モデル
- ロジスティック回帰モデル
  - 比のモデルでは対数をとって足し算モデルに

38