2019/12/9　北大・医理工統計学⑦-2

統計モデルの作り方

北海道大学　医学統計学
横田　勲

1

今回の内容

2

- 回帰モデルにおける変数選択
- 性能評価
  - 感度・特異度、ROC曲線

到達目標

- 因果モデルと予測モデルの違いを知る
- ROC曲線の意味を知る

2

医学研究での目的

3

- “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

3

因果と予測

4

- 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.

4

前立腺がんとPSA

6

- 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

6

前立腺がんのリスク因子を検討

7

- 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

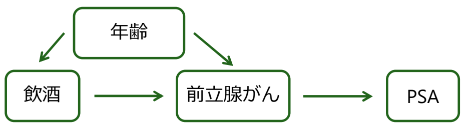
飲酒 → 前立腺がん → PSA

7

交絡の影響を解析で除去

8

- 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

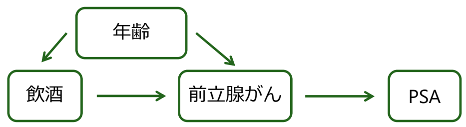


8

前立腺がんの予測をしたい

10

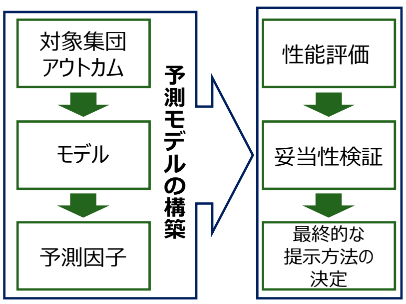
- 前立腺がん発生を精度良く当てたい
  - どのような因子を用いてもよい
    - 年齢のようなリスク因子
    - 前立腺がんの"結果"であるPSA



10

予測モデル構築の流れ

12



12

DLBCLの新規予後予測モデル

14

- びまん性大細胞型B細胞リンパ腫
- 全生存予後を予測したい
- 臨床で簡単に利用できるスコアを作りたい
  - 年齢、血清LDH、Ann Arborステージ、
    ECOG-Performance Status、血清CRP、
    低アルブミン血症、
    節外（骨髄、骨、皮膚、肺／胸膜）病変
  - 変数選択により、予測に用いる因子を決定

14

ハザード

15

- ハザード関数$\lambda(t)$

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \le T < t + \Delta t | T \ge t)}{\Delta t}$$

  - tまではat riskであるものの（$T \ge t$）、
    その直後$t + \Delta t$までにイベント発生する確率

15

Cox比例ハザードモデルを利用

16

Cox DR. JRSS B. 1972

- ハザード$\lambda(t)$に対する回帰分析

$$\lambda(t) = \lambda_0(t) \exp(x^\top \beta)$$

  - パラメータは$\lambda_0(t)$とβ
- セミパラメトリックモデル
  - 尤度関数のにβに関する部分だけ最大化
  - $\lambda_0(t)$は$\hat{\beta}$を差し込んでノンパラ推定
  - 計数過程により漸近性質が正当化

16

変数選択

17

- 回帰分析において、複数の因子候補から、関連の強そうなものだけに絞る方法
  - 予測モデルをシンプルにするためには便利
- 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

17

ランダム分割

18

- 465例のデータ
  - 323例(70%)をトレーニングコホート
  - 142例(30%)をバリデーションコホート
- トレーニングコホートで予測モデルを構築
- バリデーションコホートで他の予測モデルとの性能を比較
  - モデル構築に用いていないデータであるため、公平な性能比較を行えるだろう

18

最終モデル

19

- 変数減少ステップワイズ法を利用

因子	ハザード比	95%信頼区間	回帰係数	スコア
LDH ≤ 1×ULN	1	-	0	
LDH ＞ 1×ULN,≤3×ULN	2.47	1.20-5.08	0.91	1点
LDH ＞ 3×ULN	3.68	1.57-8.66	1.31	2点
ECOG-PS ≥ 2	2.50	1.40-4.45	0.91	1点
ALB ＜ 3.5mg/dL	2.52	1.36-4.69	0.93	1点
特定部位への節外病変	1.71	1.03-2.84	0.54	1点

- 合計点を基にさらにリスク分類

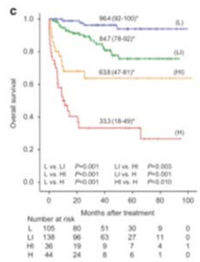
合計点	0点	1-2点	3点	4-5点
リスク分類	低	低中間	高中間	高

19

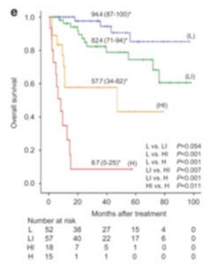
各コホートでのリスク分類

20

トレーニング



バリデーション



20

予測性能指標による評価

21

- 予後の悪い対象者を特定するための予測モデルがどれだけ有用かを知りたい
- 他の予測モデルと比較したい
- 予測モデルを構築する上で、overfittingを避けたい
  - ノイズまでモデルをあてはめてしまい、将来の対象者への予測性能が悪くなること

ところで「予測性能がよい」とはどういうこと？

21

モデルのよさ、精度の測り方

22

- 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

22

決定係数 R^2

23

- $R^2 = 1 - \frac{\underbrace{(y_i - \hat{y}_i)^2}_{残差平方和}}{\underbrace{(y_i - \bar{y})^2}_{全体の平方和}} = \frac{\overbrace{(\hat{y}_i - \bar{y})^2}^{モデルで説明した平方和}}{(y_i - \bar{y})^2}$
- 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

23

確率(密度)関数と尤度

24

- 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

24

最尤法 maximum likelihood method

25

- 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

25

尤度によるあてはまりの評価

26

- 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

26

モデルの複雑さとoverfitting

27

- 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

27

Akaike's Information Criterion; AIC

28

- $-2 \log L(\boldsymbol{\beta}; x) + 2K$
  - 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

28

古典的な変数選択法

- 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

29

他の変数選択法

- 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

30

平均二乗誤差 Mean Squared Error

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

31

Brier スコア

- 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
  - $\{I(y=1) - \hat{y}\}^2$
  - $\{I(T>t) - \hat{S}(t|\boldsymbol{X})\}^2$
    - $I(\cdot)$：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

32

平均Brierスコアの数値例

- 2人死亡、2人生存という仮想例
  - 無情報モデル

ID	生存／死亡	予測確率	Brierスコア
1	死亡	0.5	$(1-0.5)^2 = 0.25$
2	死亡	0.5	$(1-0.5)^2 = 0.25$
3	生存	0.5	$(0-0.5)^2 = 0.25$
4	生存	0.5	$(0-0.5)^2 = 0.25$

平均Brierスコア
0.25

  - 予測モデル

ID	生存／死亡	予測確率	Brierスコア
1	死亡	0.9	$(1-0.9)^2 = 0.01$
2	死亡	0.6	$(1-0.6)^2 = 0.16$
3	生存	0.3	$(0-0.3)^2 = 0.09$
4	生存	0.2	$(0-0.2)^2 = 0.04$

平均Brierスコア
0.075

33

相対Brierスコア減少

- 期待Brierスコアのとりうる範囲は0から0.25
  - しかも0に近いほど「予測性能がよい」
  - 集団全体の生存確率によって、上限が変化
- 無情報モデルに対する、予測モデルでの期待Brierスコアを小さくした割合
  - 0から1をとり、1に近いほど「予測性能がよい」

$$\frac{\text{Brier}_{無情報モデル} - \text{Brier}_{予測モデル}}{\text{Brier}_{無情報モデル}}$$

34

打ち切りを含むデータでの推定
35

- 対象者はいずれか3パターン
  1. $I(\tilde{T}_i > t)$　（tでイベント未発生）
  2. $I(\tilde{T}_i \le t)$かつ$\delta_i = 1$　（tでイベント発生）
  3. $I(\tilde{T}_i \le t)$かつ$\delta_i = 0$　（tでの状態不明）

35

IPCW法の利用
36

- 時点tにて打ち切りがない確率 $G(t)$
  - 例えばKaplan-Meier法で推定
- パターン3のBrierスコアが計算不能
  - パターン1,2のBrierスコアを
打ち切られない確率の逆数で膨らませる

$$\frac{1}{n}\left[\sum_i \frac{1}{\hat{G}(t)}\{0-\hat{S}(t|M_i)\}^2 + \sum_i \frac{1}{\hat{G}(\tilde{t}_i)}\{1-\hat{S}(t|M_i)\}^2\right]$$

全員

パターン1は
評価時点tの
打ち切りなし確率

パターン2は
各イベント時点$\tilde{t}_i$の
打ち切りなし確率

36

感度と特異度
37

	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

37

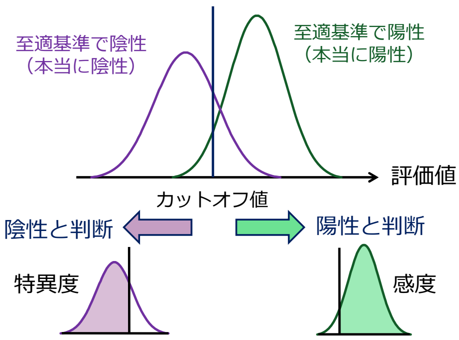
陽性的中度、陰性的中度との違い
38

	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

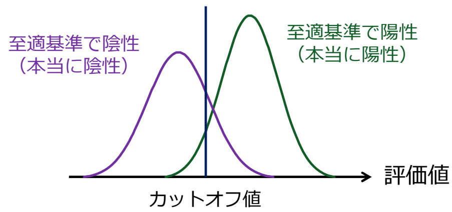
38

カットオフ値をもって評価
39



39

感度と特異度はトレードオフ
40



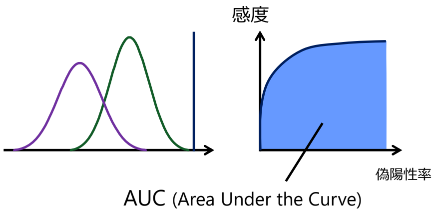
- 感度を上げれば、特異度は下がる
- 特異度を上げれば、感度は下がる

40

ROC曲線

41

- 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



41

(ROC-)AUC

42

- 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

42

c-index

43

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

- $\hat{p}_i$：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）

43

c-indexの数値例①

44

- 死亡例の予測確率：A:0.9, B:0.7, C:0.4
- 生存例の予測確率：D:0.6, E:0.3, F:0.1
- 総当たり表

		生存例		
		0.6	0.3	0.1
死亡例	0.4	×	○	○
	0.7	○	○	○
	0.9	○	○	○

○：一致　×：不一致

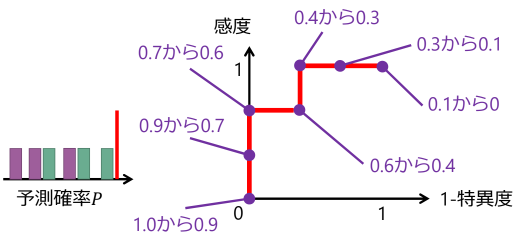
c-indexは
8/9=0.89

44

c-indexの数値例②

45

- 死亡例の予測確率：A:0.9, B:0.7, C:0.4
- 生存例の予測確率：D:0.6, E:0.3, F:0.1

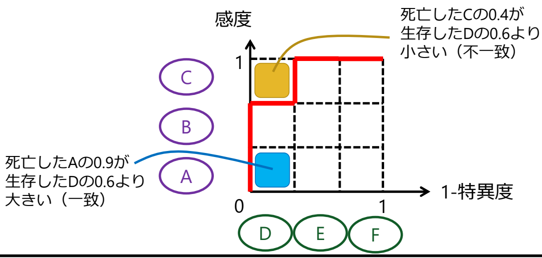


45

c-indexの数値例③

46

- 死亡例の予測確率：A:0.9, B:0.7, C:0.4
- 生存例の予測確率：D:0.6, E:0.3, F:0.1



46

c-indexを生存時間データに拡張

47

- 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.

47

DLBCL予測モデル研究

48

	PFS		OS	
	c-index	RBSR	c-index	RBSR
R-IPI	0.668	0.122	0.642	0.135
NCCN-IPI	0.749	0.172	0.736	0.251
提案スコア(4段階)	0.703	0.183	0.740	0.305
元の0-5点スコア	0.711	0.215	0.754	0.356

RBSR：相対Brierスコア減少

- 提案スコアが従来スコアより概ね性能がよいことを示した

48

練習①

49

- JMPデータを用いて、ROC曲線を描いてみよう

49

まとめ

50

- 因果モデル
  - 変数間の因果関係を仮定して、適切な条件付けを考える
- 予測モデル
  - Overfittingを防ぎながら予測性能のよいモデルを選択
- ROC曲線
  - 古典的に使われる判別性能の評価方法

50