

医学統計勉強会

第6回 ロジスティック回帰分析

帝京大学臨床研究センター（TARC）・帝京大学大学院公衆衛生学研究科 共催

帝京大学大学院公衆衛生学研究科

宮田 敏

ロジスティック回帰分析

ロジスティック回帰分析 (logistic regression analysis) は、一つのカテゴリ変数（二値変数）の成功確率を、複数の説明変数によって説明、予測する多変量解析 (multivariable analysis) の一つ。

被説明変数 Y は 0 もしくは 1 を値にとり、イベント発生の有無を表す。

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}, i = 1, \dots, n.$$

$$Y_i = \begin{cases} 1 \\ 0 \end{cases}, \quad P(Y_i = 1) = p_i.$$

Y は 0 か 1 のいずれかをとる二値変数。
 $Y = 1$ になる確率 p を、説明変数で予測したい。

Example : Risk Factors Associated with Low Infant Birth Weight

Springfield, Massachusetts にある Baystate Medical Center で収集された, 189 人の幼児のデータ. 低出生体重に対するリスクファクターを探索することが目的.

low 出生体重が2.5kgを下回るか否かのダミー変数 (0/1). **被説明変数**
age 母親の年齢 (年).
lwt 最終月経期間における母親の体重.
race 母親の人種 (1 = 白人, 2 = 黒人, 3 = その他).
smoke 妊娠期間の喫煙の有無 (0/1).
ptd 過去の早産の有無 (0/1).
ht 高血圧症の有無 (0/1).
ui 子宮炎症の有無 (0/1).
ftv 妊娠後最初の3ヶ月間に医師の診断を受けた回数. (0, 1, 2+)
(1, 2+を纏めて, (0, 1+)とした)

Hosmer, D.W. and Lemeshow, S. (1989) *Applied Logistic Regression*. New York: Wiley

Venables, W.N. and Ripley, B.D. (1999) *Modern Applied Statistics with S-PLUS*. New York: Springer-Verlag

ロジスティック回帰モデル (logistic regression model)

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}, i = 1, \dots, n.$$

$$Y_i = \begin{cases} 1 \\ 0 \end{cases}, \quad P(Y_i = 1) = p_i.$$

$\log\left(\frac{p_i}{1-p_i}\right)$: 対数オッズ (log odds), ロジット (logit)

$$\frac{p_i}{1-p_i} : \text{オッズ (odds)} \quad \log\left(\frac{p_i}{1-p_i}\right) > 0 \iff \frac{p_i}{1-p_i} > 1 \iff p_i > \frac{1}{2}$$

「対数オッズが0より大きい」 \Leftrightarrow

「オッズが1より大きい」 \Leftrightarrow 「イベントの発生確率が50%より大きい」

ロジスティック回帰モデルとイベント発生確率

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}, i = 1, \dots, n.$$

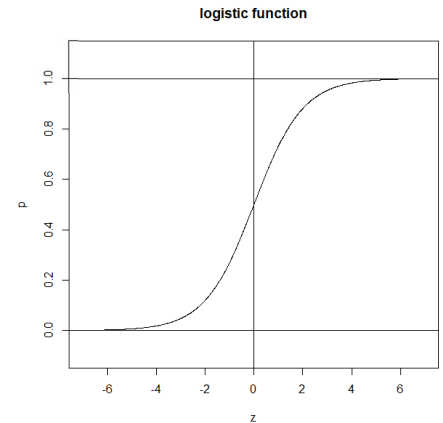
$$\Leftrightarrow p_i = \frac{\exp\{\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}\}}{1 + \exp\{\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}\}}$$

$z = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}$ とすれば、以下の

ロジスティック関数 (logistic function) が得られる

$$p = e^z / (1 + e^z) \quad \text{logistic function}$$

ロジスティック関数は、実数上で定義され 0 以上1以下の単調増加関数。



$\beta > 0 \Rightarrow x$ の増加は $z = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$ を増やす
 \Rightarrow イベント発生確率 p が上昇する

2024/11/27

医学統計勉強会 第6回 ロジスティック回帰分析

5

ロジスティック回帰モデルとオッズ比の関係

いま、 x_2, \dots, x_k が一定であったとき x_1 の値が 1 単位増加

p : 元のイベント発生確率

q : x_1 が1単位増加した後のイベント発生確率

$$\log(p/1-p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

$$\log(q/1-q) = \beta_0 + \beta_1 (x_1 + 1) + \cdots + \beta_k x_k$$

\Leftrightarrow

$$\log(q/1-q) - \log(p/1-p) = \log\left(\frac{q/1-q}{p/1-p}\right) = \beta_1$$

$$\Leftrightarrow \frac{q/1-q}{p/1-p} = e^{\beta_1} : x_1 \text{ が1単位増加した前後のオッズ比}$$

2024/11/27

医学統計勉強会 第6回 ロジスティック回帰分析

6

ロジスティック回帰モデルの推定と検定

ロジスティック回帰のパラメータは、最尤法 (MLE, Maximum Likelihood Estimation) により推定される。

Example : Low Infant Birth Weightデータ

Coefficients:	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	0.82302	1.24471	0.661	0.50848	
age	-0.03723	0.0387	-0.962	0.33602	
lwt	-0.01565	0.00708	-2.211	0.02705	*
raceblack	1.19241	0.53597	2.225	0.02609	*
raceother	0.74069	0.46174	1.604	0.10869	
smokeTRUE	0.75553	0.42502	1.778	0.07546	.
ptdTRUE	1.34376	0.48062	2.796	0.00518	**
htTRUE	1.91317	0.72074	2.654	0.00794	**
uiTRUE	0.68019	0.46434	1.465	0.14296	
ftv1	-0.43638	0.47939	-0.91	0.36268	
ftv2+	0.17901	0.45638	0.392	0.69488	

Null deviance: 234.67 on 188 degrees of freedom					
Residual deviance: 195.48 on 178 degrees of freedom					

- 回帰係数の推定値 : $\beta > 0$ であれば, x の増加はイベント確率とリスクの上昇.
- 回帰係数の有意性検定の p 値
- 推定量の標準誤差 (Std. Error信頼区間に使う)

2024/11/27

医学統計勉強会 第6回 ロジスティック回帰分析

7

ロジスティック回帰モデルの推定と検定

回帰係数の信頼区間 (CI, Confidence Interval)

CI: $\hat{\beta} \pm 1.96 \times (\hat{\beta} \text{ の標準誤差})$

$$\hat{\beta}_2 \pm 1.96s_{\hat{\beta}_2} \Leftrightarrow -0.01565 \pm 1.96(0.00708) \Leftrightarrow (-0.02953, -0.00177)$$

オッズ比の信頼区間

オッズ比 : $e^{\hat{\beta}_2} = \exp\{\hat{\beta}_2\} = \exp\{-0.01565\} = 0.9845$

信頼区間 : $(\exp\{-0.02953\}, \exp\{-0.00177\}) \Leftrightarrow (0.9709, 0.9982)$

2024/11/27

医学統計勉強会 第6回 ロジスティック回帰分析

8

ロジスティック回帰モデルの予測と判別

- ロジスティック回帰モデルの回帰係数が推定できたとする。係数の推定値を元のモデルに代入すれば、イベント発生確率の**予測式**ができる。

$$\hat{p} = \frac{\exp[\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k]}{1 + \exp[\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k]}$$

- $\hat{p} > 0.5 \Rightarrow$ イベント発生あり
- $\hat{p} < 0.5 \Rightarrow$ イベント発生なしと「判別」すれば、新しい患者さんに対してイベント発生の有無を予測できる

ロジスティック回帰モデルの適合度検定

個々の回帰係数の有意性ではなく、ロジスティック回帰モデル全体の当てはまりの良さを検定したい。

(回帰分析の model utility test に相当する)

Hosmer-Lemeshow検定 H_0 : 当てはめたモデルが正しい

イベントの予測確率に従い、標本を $k = 10$ 群に分ける。

O_i : 第 i 群のイベント発生数, N_i : 第 i 群のサンプル数

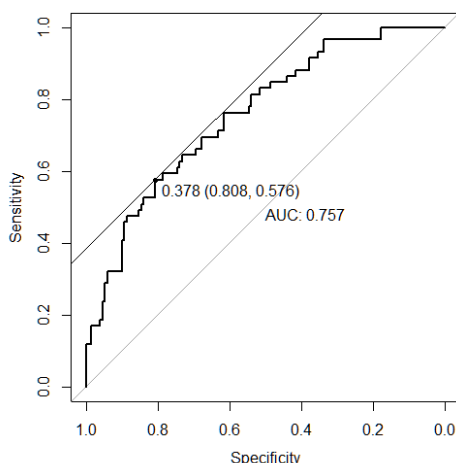
$\hat{\pi}_i$: 第 i 群の平均イベント発生確率,

検定統計量 $\chi^2 = \sum \frac{(O_i - N_i \hat{\pi}_i)^2}{N_i \hat{\pi}_i (1 - \hat{\pi}_i)} \sim \chi^2_{\text{degrees of freedom} = k - 2}$

HL検定は p 値が大きく、 H_0 を棄却できないほうが嬉しい。

ロジスティック回帰モデルの適合度(続き)

ROC曲線のAUC (Area Under the Receiver Operating Characteristic Curve) 応答変数であるイベントの有無と、ロジスティックモデルから推定された予測確率でROC曲線を描く。



2024/11/27

医学統計勉強会 第6回 ロジスティック回帰分析

11

ロジスティック回帰モデル適用の問題点

多重共線性 (multicollinearity)：説明変数の間に強い線形関係（＝比例関係）が存在する場合、推定が不安定になる。

完全分離：説明変数の値によって、イベントの発生の有無が完全に分離した場合、ロジスティック回帰の推定ができない（するまでもない）。

外れ値 (outlier)：残差 $r_i = Y_i - \hat{\pi}_i$ あるいは標準化残差 $r_i / \sqrt{\hat{\pi}_i(1 - \hat{\pi}_i)}$ で検出する。

2024/11/27

医学統計勉強会 第6回 ロジスティック回帰分析

12

変数選択

多数の説明変数の候補の中から被説明変数の変動を説明する**最適な組み合わせ**を探索する。探索の過程は、1) 変数増加法 (forward selection)
2) 変数減少法 (backward elimination)
3) 変数増減法 (stepwise procedure)

方法1：取り込む、もしくは取り除く説明変数の有意性を**逐次検定**する方法。

方法2：モデルの当てはまりの良さを測る尺度（=**モデル選択基準**）を定義し、そのモデル選択基準を最適化するように説明変数を選択する方法

変数選択（方法1）

Step1 (forward selection)：既存のモデルに説明変数を一つ加え、有意性検定のp値を求める。 P 値が「投入」確率より小さければモデルに残す。投入できる変数がなくなるまで続ける。

Step2 (backward elimination)：既存のモデルから、一つずつ説明変数を除いたときの有意性を検定しp値を求める。最も大きい p 値が「除去」確率を上回ったとき、その変数を除く。

全ての変数のp値が除去確率を下回ったとき、変数選択を止める。（投入確率、除去確率は0.1~0.2とする）

変数選択（方法2）

モデル選択基準の最適化：

AIC (Akaike's Information Criterion, **赤池の情報量基準**)

$$AIC = -2\log L + 2p$$

BIC (Bayesian Information Criterion, **ベイズ情報量基準**)

$$BIC = -2\log L + \log(n)p$$

ただし, $\log L$: 対数尤度（回帰分析における残差二乗和に当たる）, p : パラメーターの数, n : サンプル数

変数選択の実際（私の場合）

医学統計の性質上, 説明変数は取りこぼしなく広めに選択したい。しかし, 候補の数がサンプル数より大きければ, 最初に全ての変数を用いたモデルは推定できない。また, 欠測が多いとサンプル数が減る。

1. まず**単変量解析**で, 変数ごとのp値を出す。
2. ある程度p値が小さい変数に, **候補を絞る**.
($p < 0.2$ 程度で, あまりきつく絞らない)
3. 絞った候補から, **変数減少法**で選択する。(割合多めの変数が選択される) Rでは `stepAIC()` コマンドを用いる (MASSパッケージが必要)。

変数選択の問題

- 変数選択の結果選ばれたモデルに、有意でない変数が含まれる。
変数選択は、被説明変数の変動をもっともよく説明する変数の組み合わせを探索する。個々の変数が有意でなくても、組み合わせ全体として最適であると解釈する。
- 変数選択の結果、興味のある変数がモデルから除かれてしまった。
上記に矛盾するようだが、回帰の目的は被説明変数を説明することだけではない。変数間の関係を推測するため、興味ある変数を強制投入してもよい。

変数選択（Low Infant Birth Weight データ）

Coefficients:				
	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.12533	0.967561	-0.13	0.89694
lwt	-0.01592	0.006954	-2.289	0.02207 *
raceblack	1.300856	0.528484	2.461	0.01384 *
raceother	0.854414	0.440907	1.938	0.05264 .
smokeTRU	0.866582	0.404469	2.143	0.03215 *
ptdTRUE	1.128857	0.450388	2.506	0.0122 *
htTRUE	1.866895	0.707373	2.639	0.00831 **
uiTRUE	0.750649	0.458815	1.636	0.10183

Null deviance: 234.67 on 188 degrees of freedom				
Residual deviance: 197.85 on 181 degrees of freedom				

- 全ての説明変数を用いたfull modelから出発して、方法2に従いAICを最小化。
- age, ftv がモデルから脱落。
- uiは $p=0.10183$ であるが、変数選択はモデル全体のfitnessを最適化しているので、このまま残してよい。

線形モデルを超えて ー非線形モデルの世界へー

線形回帰モデルも、ロジスティック回帰モデルも、「線形性の仮定」を前提としている。

線形回帰モデル： $y_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki} + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2)$

ロジスティック回帰モデル： $\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki}$

「線形性の仮定」は、あくまで**単純化**のための仮定。現実のデータには、しばしば非線形な構造が存在する。

⇒ **非線形モデル**への、モデルの拡張。

一般化加法モデル (Generalized additive model, GAM)

線形モデルの一次式に、非線形変換を導入する。

加法モデル： $y_i = \beta_0 + f_1(x_{1i}) + \dots + f_k(x_{ki}) + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2)$

ロジスティック加法モデル $\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + f_1(x_{1i}) + \dots + f_k(x_{ki})$

f_1, \dots, f_k は x の非線形変換で、データに適合するように自動的に選ばれる。

GAMは、ソフトウェアによっては実装していないものもある。興味のある方は、ご相談ください。

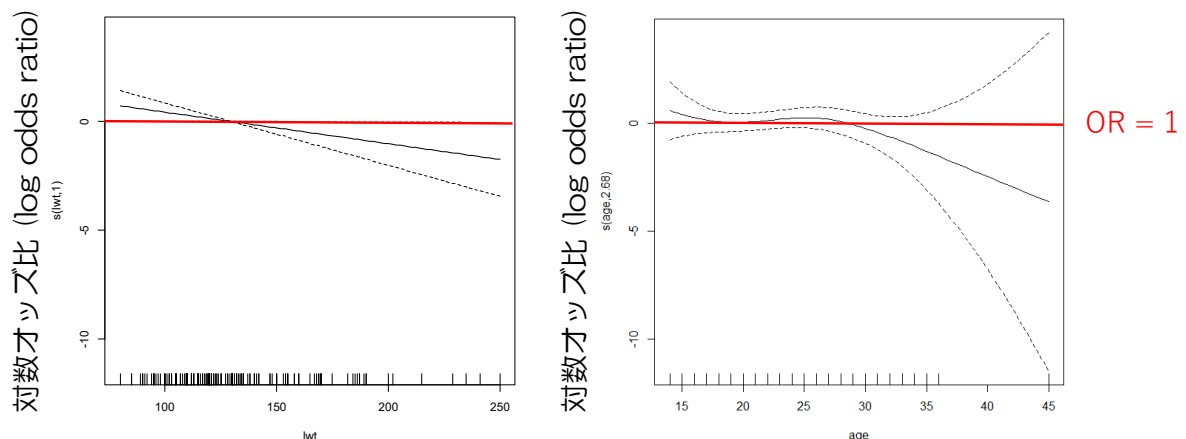
一般化加法モデル (Low Infant Birth Weight データ)

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + f_1(\text{age}) + f_2(\text{lwt}) + \text{race} + \text{smoke} + \text{ptd} + \text{ht} + \text{ui}$$

Parametric coefficients:				
	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.2253	0.4163	-5.346	9.01E-08 ***
raceblack	1.2503	0.5326	2.348	0.01889 *
raceother	0.7803	0.4502	1.733	0.08307 .
smokeTRUE	0.906	0.4118	2.2	0.0278 *
ptdTRUE	1.1749	0.4704	2.497	0.01251 *
htTRUE	1.8562	0.7109	2.611	0.00903 **
uiTRUE	0.7608	0.4694	1.621	0.10504
Approximate significance of smooth terms:				
	edf	Ref.df	Chi.sq	p-value
s(age)	2.68	3.38	3.426	0.3863
s(lwt)	1	1	4.242	0.0395 *

- 線形項については、元のロジスティックモデルと同様の結果
- 非線形項については、ageは有意ではないが、lwtは有意。

一般化加法モデル (Low Infant Birth Weight データ, 続き)



- lwtに対しては非線形な変換が選択されず。
- ageに関して30歳以前はリスクに影響を与えない一方で、30歳以降リスクが低下する傾向(有意ではない)