―基礎科学から医学・医療を見る―

ロジスティック回帰

中澤　秀夫
日本医科大学基礎科学数学

Logistic Regressions

Hideo Nakazawa
Department of Mathematics, Nippon Medical School

Abstract

This paper explains logistic regression analysis, which is a commonly used technique in medical statistics. In particular, the basic ideas and methods of interval estimation and hypothesis testing with the software package SPSS are explained.
（日本医科大学医学会雑誌　2014; 10: 186–191）

Key words: logistic regression, odds ratio, relative risk, interval estimation, hypothesis testing

1. ロジスティック回帰の原理

1.1 ロジスティック回帰とは

例えばある病気の発症の原因を追究する問題を考える. 原因と疑われる多くの因子(例えば性別や身長，体重, 血圧, コレステロール値など)のうち, どの因子がその病気の発症に真に影響を与えているかを考えたい. 今, 因子が n 個あると仮定し, これらを x_1，x_2，x_3，…，x_n と表すことにする．結果 y は発症か発症でないかのいずれかと仮定して，これを 0(発症せず)，1(発症)と表すことにする．これらのデータを各患者ごとに表にまとめたものが以下のようになっているとする：

因子 / 患者	x_1	x_2	$\cdots$	x_n	y
1	x_{11}	x_{21}	$\cdots$	x_{n1}	y_1
2	x_{12}	x_{22}	$\cdots$	x_{n2}	y_2
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
m	x_{1m}	x_{2m}	$\cdots$	x_{nm}	y_m

この問題において，各因子 x_1，x_2，x_3，…，x_n を変数とする関数で，値域が［0, 1］の値をとるものを構成したい．n 変数関数を一般的に考えるのはあまりに難しいので，線形回帰のアイデアに習い，各 b_k(k=0, 1，2，…，n）を定数として，x_1，x_2，x_3，…，x_n の線形結合である

$$X = b_0 + \sum_{k=1}^{n} b_k x_k$$

の形を仮定する．この値は因子となるデータの数値に応じて負の値になることもあるだろう．したがって $-\infty < X < +\infty$ つまり X は実数全体を変わりうる数値と思ってよい.

さて結果は 0 以上 1 以下の値をとる関数なので X に対して

$$0 \leqq y \leqq 1$$

を満たす y を対応させる関数を作りたい.

そこで次の式変形を考えよう：

$$0 < y < 1 \quad \Rightarrow \quad 0 < \frac{y}{1-y} < +\infty$$

Correspondence to Hideo Nakazawa, Department of Mathematics, Nippon Medical School, 1–7–1 Kyonancho, Musashino, Tokyo 180–0023, Japan
E-mail: nakazawa-hideo@nms.ac.jp
Journal Website（http://www.nms.ac.jp/jmanms/）

$$\Rightarrow \quad -\infty < \log\left(\frac{y}{1-y}\right) < +\infty$$

よって，この最後の式を X と思ってみると

$$X = \log\left(\frac{y}{1-y}\right) \quad \Leftrightarrow \quad e^X = \frac{y}{1-y}$$
$$\Leftrightarrow \quad e^X(1-y) = y \quad \Leftrightarrow \quad e^X = (1+e^X)y$$
$$\Leftrightarrow \quad y = \frac{e^X}{1+e^X}$$

となり，結局 n 個の変数 x_1，x_2，…，x_n に対して

$$y = \frac{\exp\left(b_0 + \sum_{k=1}^{n} b_k x_k\right)}{1 + \exp\left(b_0 + \sum_{k=1}^{n} b_k x_k\right)}$$

を考えれば良さそうである（ただし $\exp(x)$ は e^x を表す）．このような曲線による回帰をロジスティック回帰（logistic regression）[注]という．これは疾患のリスク因子を分析するためによく用いられる多変量解析手法の一つである．

変数の個数 n が1の場合を単回帰（single regression analysis），$n \geqq 2$ の場合を重回帰（multiple regression）とか多重ロジスティック回帰分析（multiple logistic regression analysis）などと呼ぶ．

各係数 b_0，b_1，b_2，…，b_n は，測定データをよく近似するように選ばれ，回帰係数（regression coefficient）という．これらは一般には最尤法（maximum-likelihood method）によって求められ，手計算ではなく SPSS などの統計ソフトを活用して求めることが多い．

1.2 オッズ

確率 p（$0 \leqq p \leqq 1$）に対して $\frac{p}{1-p}$ をオッズ（odds）という．由来は競馬における馬券のオッズで，それは，当たる確率を p とすると，

$$(\text{馬券のオッズ}) = \frac{\text{当たる確率}}{\text{当たらない確率}} = \frac{p}{1-p}$$

で与えられる．意味は

$$(\text{オッズ}) > 1 \Leftrightarrow \frac{p}{1-p} > 1 \Leftrightarrow p > 1-p$$
$$\Leftrightarrow 2p > 1 \Leftrightarrow p > \frac{1}{2} \Leftrightarrow \text{当たりやすい}$$

である．同様に，ある病気に対するオッズとは，ある病気が起こる確率を p として，その病気の起こる確率 p の，起こらない確率 $1-p$ に対する比を意味する：

$$(\text{ある病気のオッズ}) = \frac{p}{1-p}.$$

1.3 オッズ比

オッズ比（odds ratio）［OR］とは，危険因子に曝露した患者群のオッズを，危険因子に曝露していない対照群のオッズで割った値のことをいう．

例えば，コホート研究やケースコントロール研究における次のような四分表を考える（例えば曝露群を喫煙有，非曝露群を喫煙無，また患者群を肺癌有，非患者群を肺癌無と思うとよい）：

	曝露群	非曝露群	合計
アウトカム＋（患者群）	a	b	$a+b$
アウトカム－（非患者群）	c	d	$c+d$
合計	$a+c$	$b+d$	$n(=a+b+c+d)$

この表において，患者群では曝露有の確率 p は

$$p = \frac{a}{a+b},$$

したがって患者群のオッズは

$$(\text{患者群のオッズ}) = \frac{p}{1-p} = \frac{\frac{a}{a+b}}{1-\frac{a}{a+b}} = \frac{\frac{a}{a+b}}{\frac{b}{a+b}} = \frac{a}{b}.$$

同様に，非患者群で曝露有の確率 q は

$$q = \frac{c}{c+d},$$

したがって非患者群のオッズは

$$(\text{非患者群のオッズ}) = \frac{q}{1-q} = \frac{\frac{c}{c+d}}{1-\frac{c}{c+d}} = \frac{\frac{c}{c+d}}{\frac{d}{c+d}} = \frac{c}{d}$$

となる．以上よりオッズ比は

$$[\text{OR}] = \frac{\frac{p}{1-p}}{\frac{q}{1-q}} = \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{ad}{bc}$$

で与えられる．

オッズ比の意味に関しては次のとおり：

［OR］＝1⇔危険因子はアウトカムに影響を与えない

［OR］＞1⇔危険因子はアウトカムの発生を高くする

[注]1948年にアメリカのフラミンガムで開始されたフラミンガム研究（Framingham study）のために開発された．この研究は冠状動脈性疾患に関する大規模なコホート研究で，複数のリスクファクター（多重リスクファクター（multiple risk factor））が疾患に及ぼす影響を分析することを目的としたもの．

［OR］＜1⇔危険因子はアウトカムの発生を低くする

1.4 相対危険度

相対危険度（relative risk）［RR］とは，ある疾患において，曝露群での発生率と非曝露群での発生率の比のことをいう．これは，コホート研究やランダム化比較試験では算出可能である．その一方ケースコントロール研究においては，発症率や疾患の存在率（有病率）などの割合は，単に研究者が任意に決めたもので一般集団における割合を必ずしも正しく反映しているわけではないという事情のために，相対危険度を算出することはできない．

先の四分表において

$$\text{曝露群の危険度}=\frac{a}{a+c},\ \text{非曝露群の危険度}=\frac{b}{b+d}$$

であるから

$$[\mathrm{RR}]=\frac{\text{曝露群の危険度}}{\text{非曝露群の危険度}}=\frac{\dfrac{a}{a+c}}{\dfrac{b}{b+d}}=(\bigstar)$$

ここで，もしも a および b が微小量なら，$a+c\simeq c$, $b+d\simeq d$ と近似できるので

$$(\bigstar)=\frac{\dfrac{a}{c}}{\dfrac{b}{d}}=\frac{a}{c}\times\frac{d}{b}=\frac{ad}{bc}\ \cdots\cdots(\bigstar\bigstar)$$

となりオッズ比と同じ値になる．

一方，もしもこの近似が妥当でないならば

$$(\bigstar)=\frac{a}{a+c}\times\frac{b+d}{b}=\frac{a(b+d)}{b(a+c)}$$

となる．

相対危険度の意味に関しては次のとおり：

［RR］＝1⇔危険因子はアウトカムに影響を与えない

［RR］＞1⇔危険因子はアウトカムの発生を高くする

［RR］＜1⇔危険因子はアウトカムの発生を低くする

1.5 なぜオッズ比か？

医学研究においては相対リスクよりもオッズ比のほうがよく用いられる．この理由は主に以下の 3 点による：

(*i*) オッズ比は相対リスクをよく近似するから．（★★）を見よ．

(*ii*) 疫学研究において，コホート研究，ケースコントロール研究，どちらであってもオッズ比は同じ値になるから．

(*iii*) ロジスティック回帰との関連で数学的に取り扱いやすいから．

1.6 例題

ある薬剤を服用していた患者 100 人（A 群）と，服用していなかった患者 100 人（B 群）とで，その薬剤の副作用の発生割合を調べたところ，次の結果を得た：

	A群（服用有）	B群（服用無）	合計
副作用有	10	5	100
副作用無	90	95	100
合計	100	100	200

以下の各問いに答えよ．

(1) A 群の副作用発症確率 P_A を求めよ．

(2) B 群の副作用発症確率 P_B を求めよ．

(3) A 群のオッズを求めよ．

(4) B 群のオッズを求めよ．

(5) オッズ比を求めよ．

(6) 変数 x を以下のように定める：

$$x=\begin{cases}1 & (\text{服用有の場合}),\\ 0 & (\text{服用無の場合}).\end{cases}$$

この x に対して確率 $p=p(x)$ を副作用発症確率とする：

$$p=p(x)=\begin{cases}P_A & (x=1),\\ P_B & (x=0).\end{cases}$$

上の表で与えられるデータに対してロジスティック回帰分析を行うために

$$\log\left(\frac{p}{1-p}\right)=a+bx\ \cdots\cdots(*)$$

を仮定するとき，a, b の値を求め，p を x の式で表せ．なお必要なら以下の値を利用せよ．

X	2	e	3	9.5
$\log_{10}X$	0.3010	0.43429	0.4771	0.9777

解 (1) $P_A=\dfrac{10}{100}=0.1.$

(2) $P_B=\dfrac{5}{100}=0.05.$

(3) $\dfrac{P_A}{1-P_A}=\dfrac{10}{90}=\dfrac{1}{9}\simeq 0.1111.$

(4) $\dfrac{P_B}{1-P_B}=\dfrac{5}{95}=\dfrac{1}{19}\simeq 0.0526.$

(5) $[\mathrm{OR}]=\dfrac{\dfrac{1}{9}}{\dfrac{1}{19}}=\dfrac{19}{9}\simeq 2.111$

(6) ［step 1］a を求める．（$*$）において $x=0$ とすると，$p=p(x)=p(0)=P_B=0.05$ であるから

$$(*) \Leftrightarrow \log\frac{1}{19} = a \cdots (i) \Leftrightarrow a = -\log 19$$
$$= -\frac{\log_{10} 19}{\log_{10} e} = -\frac{\log_{10}(2\times 9.5)}{\log_{10} e}$$
$$= -\frac{\log_{10} 2 + \log_{10} 9.5}{\log_{10} e} = -\frac{0.3010 + 0.9777}{0.43429}$$
$$= -\frac{1.287}{0.43429} \simeq -2.9443.$$

［step 2］b を求める．（$*$）において x=1 とすると，$p=p(x)=p(1)=P_A$=0.1 であるから

$$(*) \Leftrightarrow \log\frac{1}{9} = a + b \cdots (ii)$$
$$\Leftrightarrow ((i))\text{を用いて)}\ b = \log\frac{1}{9} - \log\frac{1}{19}$$
$$= \log 19 - \log 9 = \frac{\log_{10}(2\times 9.5)}{\log_{10} e} - \frac{\log_{10} 3^2}{\log_{10} e}$$
$$= \frac{\log_{10} 2 + \log_{10} 9.5 - 2\times\log_{10} 3}{\log_{10} e}$$
$$= -\frac{0.3010 + 0.9777 - 2\times 0.4771}{0.43429}$$
$$= -\frac{1.287 - 0.9542}{0.43429} = \frac{0.3245}{0.43429} \simeq 0.7472.$$

［step 3］$p=p(x)$ を求める．（$*$）を p について解いて，

$$p = p(x) = \frac{e^{a+bx}}{1+e^{a+bx}}$$
$$= \frac{1}{e^{-a-bx}+1} = \frac{1}{e^{2.9443-0.7472x}+1}$$

を得る．

2. ロジスティック回帰による区間推定と仮説検定

SPSS を用いてどの危険因子が疾患に影響を及ぼすかをロジスティック回帰の手法によって考察する方法を説明する．まず区間推定や仮説検定についての原理を述べるが，SPSS を用いればこれらの原理による手計算を実行して判断をするという必要はなく，SPSS で与えられる数値データを見るだけで結論が得られることをあらかじめ注意しておく．

2.1 ロジスティック回帰による区間推定

n 個の因子を表す変数を x_1, x_2, …, x_n とし，$X=b_0+\sum_{k=1}^{n} b_k x_k$ と置く．SPSS によって次の数値データが得られていると仮定する：

	B	S.E.
	b_0	e_0
x_1	b_1	e_1
⋮	⋮	⋮
x_n	b_n	e_n

ここに，B はサンプルデータから得られた回帰係数（点推定値（point estimation））を，また S.E. は標準誤差（standard error）を各々表す．

信頼度 100(1−α)%（⇔危険率 100α%）での信頼区間の求め方

［step 1］次で定義される量 Er を求める：

$$Er = z(\alpha) \times (\text{S.E.}).$$

ここに，$z(\alpha)$ は標準正規分布表における両側確率 100α%点を表す．代表的な値は次のとおりである：

α	0.01	0.05	0.1
$z(\alpha)$	2.576	1.960	1.645

［step 2］求める信頼区間の両端の値は次で与えられる：

$$B \pm Er.$$

つまり母回帰係数 B_*の 100(1−α)%信頼区間は

$$B - Er \leqq B_* \leqq B + Er$$

となる．

【注意】このようにして求まる理由は以下のとおり；上の B,（S.E.）に対して母回帰係数を B_*で代表させるとき，信頼区間は

$$|z| \leqq z(\alpha) \ \text{ ただし } \ z \equiv \frac{B - B_*}{(\text{S.E.})}$$

で与えられる．この式を B_*に関して解けば上の結果を得る．

2.2 ロジスティック回帰による仮説検定

前節と同様に，次の設定のもとで考える：

n 個の因子を表す変数を x_1, x_2, …, x_n とし，$X=b_0+\sum_{k=1}^{n} b_k x_k$ と置く．SPSS によって次の数値データが得られていると仮定する：

	B	S.E.
	b_0	e_0
x_1	b_1	e_1
⋮	⋮	⋮
x_n	b_n	e_n

危険率 100α%での仮説検定（Wald 検定）

［step 1］帰無仮説「$B_*=0$」を置く．

［step 2］検定統計量 $z^2 \equiv \left(\dfrac{B}{(\mathrm{S.E.})}\right)^2$ を計算する．

［step 3］次に従って判定する：

$z^2 > \chi_1^2(\alpha) \quad \Rightarrow \quad$ 仮説を棄却する

$z^2 \leqq \chi_1^2(\alpha) \quad \Rightarrow \quad$ 仮説を棄却しない

【注意】(*i*) Wald 統計量とは理論と実際のずれを評価する量であり，

$$z = \frac{B - B_*}{(\mathrm{S.E.})}$$

で与えられる．これは近似的に標準正規分布に従うことが知られている．したがって，帰無仮説 $B_*=0$ のもとでは，

$$z = \frac{B}{(\mathrm{S.E.})}$$

であるが，z^2 は自由度 1 の χ^2 分布に従うことが知られている．これについては χ^2 分布の一般論である次の事実による：

> 定理　X が標準正規分布 $N(0,1^2)$ に従うとき，X^2 は χ_1^2 に従う．

(*ii*) Wald 検定（Wald test）とは，Abraham Wald（1902〜1950）によって，

Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large, Transactions of the American Mathematical Society, 54,（1943），pp. 426–482

において導入された検定法．なお，Wald 検定という用語が初めて使われた論文は

S.D. Silvey, The Lagrangian Multiplier Test, Annals of Mathematical Statistics, 30,（1959），pp. 389–407

である．

(*iii*) Wald 統計量の意味は次のとおり：

Wald 統計量の値が大 ⇒ 仮説を棄却

Wald 統計量の値が小 ⇒ 仮説を棄却せず

2.3　例題

下の表 A は，新生児 10 人の出生体重と後遺症の有無に関する（仮想的）データである．なお表中の y_n は次で定義される：

$$y_n\,(\text{後遺症の有無}) = \begin{cases} 0 & (\text{後遺症無}) \\ 1 & (\text{後遺症有}) \end{cases}$$

表 A のデータをロジスティック回帰の方法によって SPSS を用いて解析したところ，母回帰係数に関して以下の結果 B が得られた：

表 A

患者番号 (n)	体重 (x_n)	後遺症 (y_n)
1	2.0	1
2	2.2	0
3	2.3	1
4	2.7	0
5	2.8	0
6	2.9	0
7	3.0	0
8	3.1	0
9	3.2	0
10	3.3	0

結果 B

	B	S.E.
a	17.718	14.445
b	−7.825	6.367

ただし $p = \dfrac{e^X}{1+e^X}, \quad X = a + bx.$

(1) 母回帰係数 b_* の 95% 信頼区間を求めよ．

(2) 出生体重は，後遺症発生の危険因子であると言えるか？危険率 5% で検定せよ．

解 (1)［step 1］$\alpha=0.05$ であるから $z(\alpha)=1.960$，したがって Er は

$$Er = 1.96 \times 6.367 = 12.47932.$$

［step 2］求める信頼区間は

$$-7.825 - 12.47932 \leqq b_* \leqq -7.825 + 12.47932$$

$$\Leftrightarrow \ -20.30432 \leqq b_* \leqq 4.65432$$

となる．

(2)［step 1］帰無仮説 $b_*=0$ を置く．したがって

$$z = \frac{b}{\mathrm{S.E.}} = -\frac{7.825}{6.367}$$

である．

［step 2］検定統計量は

$$z^2 = \left(\frac{b}{\mathrm{S.E.}}\right)^2 = \left(-\frac{7.825}{6.367}\right)^2 = \frac{61.230625}{40.528689} \simeq 1.51.$$

［step 3］χ^2 分布の数表によって $\chi_1^2(0.05)=3.84$ であるから $z^2 > \chi_1^2(0.05)$ が成り立たない．したがって帰無仮説 $b_*=0$ は棄却されず，出生体重が後遺症発生の危険因子とは言えない．

【注意】(1) によって，信頼区間に 0 が含まれるので $b_*=0$ の可能性もある．したがって帰無仮説を棄却できないことが (1) の時点でも分かる．

2.4 SPSS によるロジスティック回帰分析

2.3 節における例題のデータを SPSS で分析すると次のような数値データが得られる：

B	Wald	自由度	p-value
-7.825	1.511	1	0.219

この表の意味は次のとおり：

点推定値 b が -7.825 であり，標準誤差 S.E による Wald 統計量の値が 1.511（2.3 節の例題の 解 の(2) の［step 2］の計算式を見よ），これは自由度 1 の χ^2 分布に従い，Wald 統計量の値が 1.511 を超える確率が 0.219 である：

$$P(\text{Wald} > 1.511) = 0.219.$$

なお，p 値（p-value＝probability value）とは帰無仮説が正しいという条件のもとで，検定統計量の値より大きな値が得られる確率を表す．

一般に，SPSS の結果が

B	Wald	自由度	p-value
b	z^2	1	p

となった場合に，危険率 α での帰無仮説 $b_*=0$ の仮説検定においては次が成り立つ：

$\alpha > p \quad \Rightarrow \quad$ 仮説は棄却される．

$\alpha \leqq p \quad \Rightarrow \quad$ 仮説は棄却されない．

上の例題のデータにおいては，危険率 $\alpha=0.05$ とすると，$p=0.219$ に対して $\alpha \leqq p$ が成り立つので仮説は棄却されず，出生体重が後遺症発生の危険因子とは言えない．

2.5 例題

患者 26 人に対して，手術直後に実施したある検査の数値，性別，術後 1 週間以内に発症した合併症の有無を調べたデータを SPSS で分析したところ，以下のようになった：

B	Wald	自由度	p-value
-0.332	4.633	1	0.031
1.084	1.299	1	0.254

危険率 5% で検定をする場合，検査数値，性別のどちらが合併症の危険因子となりうるか？

解 検査数値に関しては，$\alpha=0.05>0.031=p$ が成り立つので仮説は棄却され，検査数値は合併症の危険因子と言える．性別に関しては，$\alpha=0.05 \leqq 0.254=p$ が成り立つので仮説は棄却されず，性別は合併症の危険因子とは言えない．

3. 文献ガイド

統計学の基礎的内容については，儀我[5]（より簡潔な儀我[3,4] も見よ）や加納–高橋[2] を参照のこと．オッズ比については加納–高橋[2] の 2 章，森實[13] の 2 章，ロジスティック回帰に関しては加納–高橋[2] の 10 章，浜田[11] の 6 章，森實[13] の 7 章，Matthews et. al.[12] の 11 章，丹後[6] の 13 章，丹後–山岡–高木[7] を，SPSS との関連については，石村–謝–久保田[1] の 4 章，対馬[8] の 14 章，対馬[9] の 5，6 章などを見よ．なお本論説は，これらの文献を参考に作成した 2013 年度日本医科大学臨床系大学院講義の講義録[10] に基づいている．

文 献

1. 石村貞夫，謝 承泰，久保田基夫：SPSS による医学・歯学・薬学のための統計解析（第 3 版），2011，東京図書.
2. 加納克己，高橋秀人：基礎医学統計学(改訂第 6 版)，2011，南江堂.
3. 儀我真理子：正規母集団，推定の考え方．日本医科大学医学会雑誌 2014; 10: 16–20.
4. 儀我真理子：検定の考え方，独立性の検定，日本医科大学医学会雑誌 2014; 10: 115–119.
5. 儀我真理子：確率・統計の基礎. 2014, ムイスリ出版.
6. 丹後俊郎：医学への統計学（第 3 版）(統計ライブラリー)，2013，朝倉書店.
7. 丹後俊郎，山岡和枝，高木晴良：新版ロジスティック回帰分析，(統計ライブラリー)，2013，朝倉書店.
8. 対馬栄輝：SPSS で学ぶ医療系データ解析―分析内容の理解と手順解説．バランスのとれた医療統計入門，2007，東京図書.
9. 対馬栄輝：SPSS で学ぶ医療系多変量データ解析―分析内容の理解と手順解説．バランスのとれた医療統計入門，2008，東京図書.
10. 中澤秀夫：医学・医療統計学入門，2013 年度日本医科大学臨床系大学院講義配布プリント，2013.
11. 浜田知久馬：学会・論文発表のための統計学―統計パッケージを誤用しないために．2012，真興交易医書出版部.
12. Matthews D. E., Farewell V. T.：実践医学統計学,（宮原英夫，折笠秀樹監訳，小田英世，寺良向聡，森田智視訳），2005，朝倉書店.
13. 森實敏夫：入門医療統計学―Evidence を見出すために，2004，東京図書.

（受付：2014 年 4 月 22 日）
（受理：2014 年 6 月 4 日）