Logistic回帰分析の分散を考える

株式会社バイオスタティスティカル　リサーチ
古川敏仁

内容

1. 二項分布とロジット（logit）g(x)= ln（p/(1-p)）の分散の違い..... 1
1.1. デルタ法によるロジット=対数オッズの分散推定..... 1
1.2. 二項分布最尤法によるロジット=対数オッズの分散推定..... 2
1.3. デルタ法および二項分布最尤法によるロジット=対数オッズの分散推定のまとめ 3
2. ロジスティック回帰説明変数の分散..... 3
2.1. ロジスティック回帰説明変数の分散..... 3
2.2. もし、ロジット変換ではないリンク関数を使った場合..... 4
3. 補足資料１　デルタ法..... 7
4. 補足資料２　予測精度確認のためのSASプログラム..... 7

1. 二項分布とロジット（logit）g(x)= ln（p/(1-p)）の分散の違い

Logistic回帰分析を理解するうえでロジット（logit）g(x)= ln（p/(1-p)）の分散を理解することは重要である。

Var｛p｝=pq/n　である。

では、Var｛ln[p/(1-p)]｝はどうなっているのであろうか？

1.1. デルタ法によるロジット=対数オッズの分散推定

例えば、デルタ法でVar｛ln[p/(1-p)]｝を求めてみよう。デルタ法というのは、確率変数Xが平均mと分散σ^2が既知の場合、その関数f(x)のx=mの近傍における分散は漸近的に(1)式で示されると言うものである。

$$Var\{f(x)\}=\{f''(m)\}^2*Var(x) \qquad (1)$$

f’(p)= logit=ln(p/(1-p))とし、p=p^の周りの分散を(1)より求めると

$$var(f'(p))=\{f'(p^)\}^2 var(p^) \qquad (2)$$

ここで、

$$f'(p)=\{ln(p/(1-p))\}'=\{ln(p)-ln(1-p))\}'=1/p+1/(1-p)=(p+1-p)/(p(1-p))=1/(pq) \qquad (3)$$

ゆえに、

var(f'(p^))= { 1/(p^q^)}^{2}var(p^)={ 1/(p^q^)}^{2}×p^q^/n=1/np^q^

つまり、ロジット（logit）g(x)= ln（p/(1-p)）＝対数オッズの分散はvar(g(p))= 1/npqとなり、また、n回の試行における　確率p=n1/nとすれば（n1：イベント数、n0：非イベント数）、ロジット＝対数オッズの分散は以下となる。

$$Var(\log(odds)) = \frac{1}{npq} = \frac{1}{np} + \frac{1}{nq} = \frac{1}{n1} + \frac{1}{n2} \quad (4)$$

ちなみに、2×2分割表で、曝露のイベント数a、非イベント数b、非曝露のイベント数c、非イベント数dのオッズ比の対数の分散も同様に以下に求められる。

$$Var\left\{\log\left(\frac{\frac{a}{b}}{\frac{c}{d}}\right)\right\} = Var\left\{\log\left(\frac{a}{b}\right) - \log\left(\frac{c}{d}\right)\right\}$$

ここで、$\log\left(\frac{a}{b}\right)$と$\log\left(\frac{c}{d}\right)$は独立だから、

$$Var\left\{\log\left(\frac{a}{b}\right) - \log\left(\frac{c}{d}\right)\right\} = Var\left\{\log\left(\frac{a}{b}\right)\right\} + Var\left\{\log\left(\frac{c}{d}\right)\right\} = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$$

1.2.　二項分布最尤法によるロジット=対数オッズの分散推定

次は、Var｛ln[p/(1-p)]｝をロジスティック回帰の最尤法から求めてみよう。ln[p/(1-p)]= β 0というモデルを考える。最尤法のパラメータ推定値の分散はRao(1973)の定理により対数尤度関数の2回微分の負値行列の逆行列で求められるから、まず、このモデルのパラメータ β 0の2回微分を求めると(6)となる。

$$L(\boldsymbol{\beta}) = \ln\{l(\boldsymbol{\beta})\} = \sum_{i=1}^{n}\left\langle yi\,\ln(\pi(xi)) + (1-yi)\ln(1-\pi(xi))\right\rangle \quad (1.4)$$

$\pi(xi) = \frac{\exp(\beta 0)}{1+\exp(\beta 0)}$を上記に代入すると

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left\langle yi\,\ln(\frac{\exp(\beta 0)}{1+\exp(\beta 0)}) + (1-yi)\ln(1-\frac{\exp(\beta 0)}{1+\exp(\beta 0)})\right\rangle$$

$$= \sum_{i=1}^{n}\left\langle yi\beta 0 - \ln(1+\exp(\beta 0))\right\rangle$$

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta 0} = \sum_{i=1}^{n}\left\langle yi - \frac{\exp(\beta 0)}{1+\exp(\beta 0)}\right\rangle = \sum_{i=1}^{n}\left\langle yi - \pi(xi)\right\rangle \quad (5)$$

$$\frac{\partial^2 L(\boldsymbol{\beta})}{\partial\beta 0\partial\beta 0}=\sum_{i=1}^{n}\left\langle yi-\frac{\exp(\beta 0)}{1+\exp(\beta 0)}\right\rangle'=-\left(\sum_{i=1}^{n}\left\langle\frac{\exp(\beta 0)}{1+\exp(\beta 0)}\right\rangle\right)'=-\left(\sum_{i=1}^{n}\left\langle 1-\frac{1}{1+\exp(\beta 0)}\right\rangle\right)'=$$

$$=\left(\sum_{i=1}^{n}-\left\langle\frac{1}{1+\exp(\beta 0)}\right\rangle^{2}\exp(\beta 0)'\right)=\left(\sum_{i=1}^{n}-\left\langle\frac{1}{1+\exp(\beta 0)}\right\rangle^{2}\exp(\beta 0)\right)$$

$$=-\sum_{i=1}^{n}\left(\frac{\exp(\beta 0)}{1+\exp(\beta 0)}\right)\left(\frac{1}{1+\exp(\beta 0)}\right)=-\sum_{i=1}^{n}\left(\frac{\exp(\beta 0)}{1+\exp(\beta 0)}\right)\left(1-\frac{\exp(\beta 0)}{1+\exp(\beta 0)}\right)$$

$$=-\sum_{i=1}^{n}\pi i(1-\pi i)\quad(6)$$

すると、最尤法からロジット＝対数オッズの分散(7)式が得られる。

$$I(\beta 0)=\sum_{i=1}^{n}\pi i(1-\pi i)$$

$$Var(\beta 0)=I^{-1}(\beta 0)=\frac{1}{\sum_{i=1}^{n}\pi i(1-\pi i)}$$

ここで、$\pi i=\hat{p}$とすれば$Var(\beta 0)=\frac{1}{n\hat{p}\hat{q}}$　(7)

1.3.　デルタ法および二項分布最尤法によるロジット=対数オッズの分散推定のまとめ

それでは、以上の結果から何がわかるのであろうか。

(1)　Var｛p｝=pq/n　Var｛ln[p/(1-p)]｝=1/npq

つまり、確率pの場合は、p=0.5、q=0.5の場合分散がもっとも大きいのに対して、ロジット=対数オッズの場合は逆に、p=0.5、q=0.5、オッズ=1.0の場合分散が最も小さいということである。

(2)　また、2×2の分割表の対数オッズ比の分散　$Var\{\log(oddsrario)\}=\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}$

などは、デルタ法から求めていることがわかる。デルタ法は漸近的な手法なので、a、b、c、dがある程度の数以上でないと、この分散の理論的な推定精度がわるいことがわかる。

(3)　デルタ法および二項分布最尤法によるロジット=対数オッズの分散推定が同じ結果になるということは重要で、われわれが、新たな手法に出会ったとき、その手法が妥当であるか常に確認する必要があるが、このように2方法を使って同様な結果が導けることを確認する習慣は重要である。

2. ロジスティック回帰説明変数の分散

2.1.　ロジスティック回帰説明変数の分散

ロジスティック回帰係数の分散は、page34-35 より、

$$\hat{\mathrm{var}}\left[\hat{\beta}\right]=\hat{\mathrm{I}}^{-1}(\hat{\beta})=(\mathrm{X}'\mathrm{V}\mathrm{X})^{-1}$$

$$X'VX = \begin{bmatrix} 1 & 1 & \cdots 1 \\ x_1 & x_2 & \cdots x_n \end{bmatrix} \begin{bmatrix} \hat{\pi}_1(1-\hat{\pi}_1) & & & \\ & \hat{\pi}_2(1-\hat{\pi}_2) & & \\ & & \ddots & \\ & & & \hat{\pi}_n(1-\hat{\pi}_n) \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

$$= \begin{bmatrix} \sum \hat{\pi}_i(1-\hat{\pi}i_i) & \sum x_i\hat{\pi}_i(1-\hat{\pi}_i) \\ \sum x_i\hat{\pi}_i(1-\hat{\pi}_i) & \sum x_i{}^2\hat{\pi}_i(1-\hat{\pi}_i) \end{bmatrix}$$

$\hat{\mathrm{var}}\left[\hat{\beta}\right] = (X'VX)^{-1}$　　なので、$\beta 1$の分散は、

$$Var(\beta 1) = \frac{\sum \hat{\pi}_i(1-\hat{\pi}_i)}{\sum x_i{}^2\hat{\pi}_i(1-\hat{\pi}_i)\sum \hat{\pi}_i(1-\hat{\pi}_i) - \sum x_i\ \hat{\pi}_i(1-\hat{\pi}_i)\sum x_i\ \hat{\pi}_i(1-\hat{\pi}_i)} \quad (8)$$

となる。

かりに、$\hat{\pi}_i(1-\hat{\pi}_i)$がiによらず一定で$\hat{\pi}_0(1-\hat{\pi}_0)$であるならば(8)式は、

$$\frac{1}{\hat{\pi}_0(1-\hat{\pi}_0)\left(\sum x_i{}^2 - \sum x_i\ \times \sum x_i\right)} = \frac{1}{\hat{\pi}_0(1-\hat{\pi}_0)\sum (x_i - \bar{x})^2}$$

となる。

つまり、$\beta 1$が安定的に推定されるためには、説明変数の値はxの平均近辺から離れる、あるいは、πiの推定値（条件付平均）が0.5に近いものが多いことである。これは、1.3　(1)と共通する事項である。

2.2.　もし、ロジット変換ではないリンク関数を使った場合

一般化線形モデル　SAS　GENMOD プロシジャでは、二項分布の回帰モデルはロジスティック（link=logit）ばかりでなく、比例確率(link=id）も推定可能である。このようなモデルではパラメータの誤差構造はどのようになってしまうのか興味があるし、実務上、重要なところである。

ロジット　ロジスティックモデル　$g(p)=\ln[p/(1-p)] = \beta 0 + \beta 1$

ID　　比例確率モデル　　$g(p)=p= \beta 0 + \beta 1$

比例確率モデルの最尤法によるパラメータ推定値の分散を Rao(1973)の定理により求めてみると以下になる。

$$L(\boldsymbol{\beta}) = \ln\{l(\boldsymbol{\beta})\} = \sum_{i=1}^{n}\left\langle yi\,\ln(\pi(xi)) + (1-yi)\ln(1-\pi(xi))\right\rangle \quad (1.4)$$

$\pi(xi) = \beta 0 + \beta 1 xi$を上記に代入すると

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left\langle yi\,\ln(\beta 0 + \beta 1 xi) + (1-yi)\ln(1-\beta 0 - \beta 1 xi)\right\rangle$$

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta 0} = \sum_{i=1}^{n}\left\langle yi\frac{1}{\beta 0 + \beta 1 xi} - (1-yi)\frac{1}{1-\beta 0 - \beta 1 xi}\right\rangle$$

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta 1} = \sum_{i=1}^{n}\left\langle yi\frac{xi}{\beta 0 + \beta 1 xi} - (1-yi)\frac{xi}{1-\beta 0 - \beta 1 xi}\right\rangle$$

$$\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta 0 \beta 0} = \sum_{i=1}^{n}\left\langle yi\frac{-1}{(\beta 0 + \beta 1 xi)^2} - (1-yi)\frac{1}{(1-\beta 0 - \beta 1 xi)^2}\right\rangle$$

$$\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta 1 \beta 1} = \sum_{i=1}^{n}\left\langle yi\frac{xi^2}{(\beta 0 + \beta 1 xi)^2} - (1-yi)\frac{xi^2}{(1-\beta 0 - \beta 1 xi)^2}\right\rangle$$

今　n回試行中、n1個のイベントが観察された確率の分散を比例確率モデルから求めてみると。モデルは

　$\beta 0$=p^=n1/n　であり、Var($\beta 0$）=pq/n　となり、二項分布に基づくpの分散に一致する(9)。

$$I(\beta 0) = -\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta 0 \beta 0} = -\sum_{i=1}^{n}\left\langle yi\frac{-1}{(\beta 0)^2} - (1-yi)\frac{1}{(1-\beta 0)^2}\right\rangle$$

$$= \frac{n1}{(\beta 0)^2} + \frac{n0}{(1-\beta 0)^2} = \frac{n}{\hat{p}} + \frac{n}{\hat{q}} = n\frac{\hat{p}+\hat{q}}{\hat{p}\hat{q}} = n\frac{1}{\hat{p}\hat{q}}$$

$$Var(\beta 0) = \frac{\hat{p}\hat{q}}{n} \quad (9)$$

　$\beta 1$の分散は、ロジスティック回帰のようなきれいな形で導出できないが、$\beta 0$の分散からある程度は推定できる。それは、$\beta 1$が安定的に推定されるためには、説明変数の値はxの平均近辺から離れる、あるいは、πiの推定値（条件付平均）が0.5から遠いいものが多いことである。

　つまり、$\beta 1$を安定的に推定するためにはπiに関しては、以下のことが示される。

　ロジスティックモデル　：できるだけ0.5に近いものが多いほうが良い

　比例確率モデル　　：できるだけ0.5から遠いものが多いほうが良い

それでは、これらの事項が実際にどの程度パラメータ推定値に影響するのかをCHDのデータを使って確認してみよう。

　CHDのデータは、年齢に伴うCHD発症リスクが0から1.0まで分布し、各年齢層で例数が均等に存在するデータである（図1）。ロジスティック回帰の年齢の条件付Logit推定値(図2)、比例確率モデルの年齢の条件付確率予測値(図4)とも年齢平均44.4歳近辺で最も推定精度がよく,年齢平均から離れるに連れ推定精度が低下していることがわかる。ロジスティックモ

デルではロジットを確率予測値に変換すると、Logit変換の特性から、確率予測値の推定精度は年齢平均から離れてもそれほど低下しないことがわかる（図3）。逆に、年齢平均近辺では、多少、ロジスティック回帰よりも比例確率モデルの方が推定精度良いことがわかる。

それでは、一般的に、ロジスティックモデル、比例確率モデルどちらを利用したほうが良いのか。

一般的には、予測確率0から1.0近辺まで精度良く推定できるロジスティックの方が安心である。しかし、結果の解釈が治療効果のOdds比ではなく、治癒確率の差である場合や、狭い範囲の条件付確率が問題になる場合は、比例確率モデルの方が有効な場合も多い。

ただ、いずれにしろデータ特性からモデルの推定精度を事前に予測、検討しておくことが重要である。

図1　CHDデータのイベント発生率

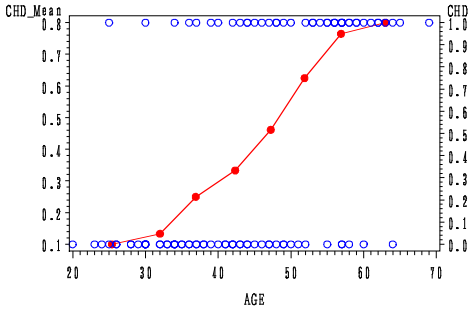


図2　年齢とLogit推定値、95%信頼区間

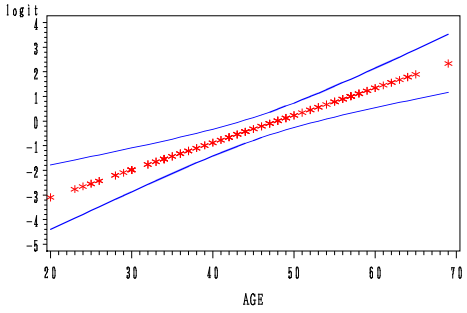


図3　Logistic予測精度＝95%信頼区間

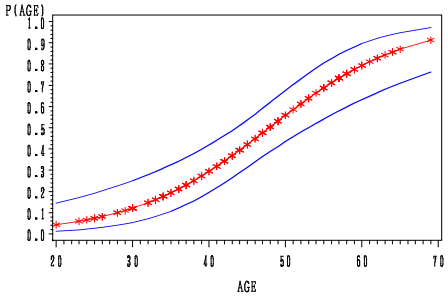
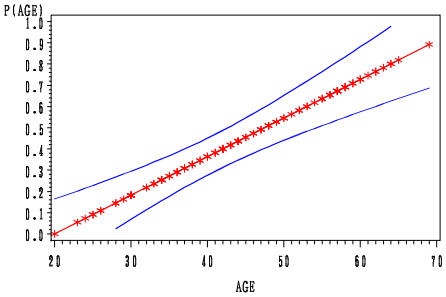


図4　比例確率予測精度=95%信頼区間



3. 補足資料 1　デルタ法

確率変数Xが平均mと分散σ^2が既知の場合、その関数f(x)のx=mの近傍の分散は漸近的に(1)式で示されると言うものである。

$$Var\{f(x|x=m)\}=\{f''(m)\}^2*Var(x) \qquad (1)$$

今仮にf(x)がn階微分可能であるとすると、f(x)は平均値mの周りでテーラー展開（Taylor Series Expansion）によりxの多項式であらわすことができる。

$$f(x) = f(m) + f'(m)(x-m) + f''(m)(x-m)^2/2+・・・・$$

ゆえに、一次までの近似式で考えると

$$f(x) \fallingdotseq f(m) + f'(m)(x-m)$$

すると、f(x)のx=mの近傍の分散は

$$Var\{f(x)\} \fallingdotseq Var\{f(m)\} + Var\{f'(m)(x-m)\} = 0 + \{f''(m)\}^2 * Var(x)$$

4. 補足資料 2　予測精度確認のためのSASプログラム

```
*　ロジスティックモデルによる CHD 発症予測
PROC LOGISTIC　OUTEST=outt；
 MODEL CHD(EVENT='1')=AGE;
 output out=out p=p lcl=lcl ucl=ucl XBETA=XBETA STDXBETA=std;
RUN;

SYMBOL1 V=star C=RED;　SYMBOL2 V=none C=BLUE I=JOIN W=2;
SYMBOL3 V=none C=BLUE I=JOIN W=2;
 *図 2 作図　　;
data outb;　　set out;
 k=1;logit=XBETA;output;　　k=2;logit=XBETA-1.96*std;output;
 k=3; ;logit=XBETA+1.96*std;output;　　run;
PROC GPLOT DATA=outb;
 PLOT logit*AGE=k/nolegend;
RUN;
 *図 3 作図
data outa;　set out;
```

```
 k=1;pp=p;output;          k=2;pp=lcl;output;          k=3;pp=ucl;output;
run;
PROC GPLOT DATA=outa;
 PLOT pp*AGE=k/nolegend;
RUN;
*   比例確率モデルによる CHD 発症予測
proc genmod data=d.anl;
 MODEL CHD=AGE /link=id dist=bin;
 output out=out p=p l=l u=u;
run;
  *図 4 作図
Data outa;       set out;
 k=1;pp=1-p;output;   k=2;pp=1-l;output;    k=3;pp=1-u;output;
run;
PROC GPLOT DATA=outa;
 PLOT pp*AGE=k/nolegend;
 LABEL PP='P(AGE)';
RUN;
```