医学統計勉強会

東北大学病院循環器内科・東北大学臨床研究推進センター　共催

東北大学大学院医学系研究科 EBM 開発学寄附講座

宮田　敏

“Data! data! data!” he cried impatiently. “I can't make bricks without clay.”
From *The Adventure of the Copper Beeches, The Adventure of Sherlock Holmes.*

「データ！データ！データ！」ホームズはいらいらして叫んだ。「粘土が無ければレンガは作れない」

第 3 回　ロジスティック回帰分析

1. ロジスティック回帰モデル

第 2 回で取り上げた回帰分析は、一つの連続数（実数）の値を複数の変数によって説明、予測する多変量解析モデルの一つでした。しかし問題によっては、予測したい**被説明変数が離散のカテゴリ変数**である場合もあります。例えば「ある薬剤を投与したとき副作用が起こるか否か」、「妊娠中喫煙した場合、低体重児が生まれるか否か」といった、起こりうる場合が 0 か 1，あるいは Yes か No に二分される（dichotomous）場合などが、カテゴリ変数が被説明変数になる最も基本的な場合です。今回取り上げる**ロジスティック回帰モデル（logistic regression model）**とは、このようなカテゴリ変数に対して、イベントが起こる確率（=成功確率）を複数の変数によって説明，予測する多変量解析モデルになります。

いま、$i=1,\ldots,n$ のサンプルに対して、被説明変数 Y_i はイベントが起こったとき $Y_i=1$，イベントが起こらなかったとき $Y_i=0$ をとる二値変数であるとします。イベントは起こる可能性も起こらない可能性もありますが、イベントが起こる確率（=成功確率）を $P(Y_i=1)=p_i$ とします。ここで注意したいのは、被説明変数 Y_i にも成功確率 p_i にもサンプル番号を表す添え字 i が付されていることです。つまり、サンプルが異なればイベントが起こる確率も当然異なるということです。このサンプルごとに異なる成功確率 p_i を、説明変数 $x_1,\ldots,x_n$ の情報を使って予測、説明しようというのが**ロジスティック回帰モデル**であると言えます。

いま、成功確率 p_i を説明変数 $x_1,\ldots,x_n$ で説明しようとしたとき、最も単純に考えれば、p_i を被説明変数とした回帰モデルを適用することが考えられます。

$$p_i=\beta_0+\beta_1 x_{1i}+\cdots+\beta_k x_{ki}+\varepsilon_i, i=1,\ldots,n$$

しかしこのモデルには，重大な欠点があります。なぜなら，左辺の p_i は確率で

すから 0 から 1 までの値しかとらないのに対して、右辺は説明変数 x の値に従って任意の実数値を取り得るからです。

ロジスティック回帰モデルでは、被説明変数 Y_i の成功確率 p_i と説明変数 $x_1,\ldots,x_n$ の間に，次のような関係を想定します。

ロジスティック回帰モデル：

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

上の式の左辺 $\log\left(\frac{p_i}{1-p_i}\right)$ は、成功確率 p_i の**ロジット (logit)、対数オッズ (log odds)** と呼ばれるものです。ロジットに含まれるオッズ (odds) $\frac{p_i}{1-p_i}$ は，次のような性質を持っています。

$$\frac{p_i}{1-p_i}>1 \Leftrightarrow p_i>1-p_i \Leftrightarrow p_i+p_i>1 \Leftrightarrow p_i>0.5$$

すなわち「オッズが 1 より大きい」ことは「イベントの発生確率が 50%より大きい」ことと同値であって、オッズはイベント発生のリスク指標の一つになっています。

ロジスティック回帰とオッズの関係：

ロジスティック回帰モデルは、 p_i のロジット（オッズの対数変換）を説明変数 $x_1,\ldots,x_n$ の一次式で表すものですが、対数変換が連続で単調増加な関数であることから、「係数 $\beta>0$ ならば、x の増加はオッズを増やす」、すなわち「正の係数 β は、x 上昇に伴うリスクの増加を意味する」と解釈することが出来ます。

ロジスティック回帰とイベント発生確率の関係：

さらにロジスティック回帰モデルの係数 β を解釈するために、ロジットを分解

してイベントの発生確率 p_i を明示的に示すと、以下のように変形することが出来ます。

$$\log\left(\frac{p_i}{1-p_i}\right)=\beta_0+\beta_1x_{1i}+\cdots+\beta_kx_{ki} \Leftrightarrow \frac{p_i}{1-p_i}=\exp\left[\beta_0+\beta_1x_{1i}+\cdots+\beta_kx_{ki}\right]$$
$$\Leftrightarrow p_i=(1-p_i)\exp\left[\beta_0+\beta_1x_{1i}+\cdots+\beta_kx_{ki}\right]$$
$$\Leftrightarrow p_i+p_i\exp\left[\beta_0+\beta_1x_{1i}+\cdots+\beta_kx_{ki}\right]=\exp\left[\beta_0+\beta_1x_{1i}+\cdots+\beta_kx_{ki}\right]$$
$$\Leftrightarrow p_i\left(1+\exp\left[\beta_0+\beta_1x_{1i}+\cdots+\beta_kx_{ki}\right]\right)=\exp\left[\beta_0+\beta_1x_{1i}+\cdots+\beta_kx_{ki}\right]$$
$$\Leftrightarrow p_i=\frac{\exp\left[\beta_0+\beta_1x_{1i}+\cdots+\beta_kx_{ki}\right]}{1+\exp\left[\beta_0+\beta_1x_{1i}+\cdots+\beta_kx_{ki}\right]}$$

最後の式で、$z=\beta_0+\beta_1x_1+\cdots+\beta_kx_k$ と置くと

$$p=\frac{\exp[z]}{1+\exp[z]}$$

と書くことが出来、これを z のロジスティック関数と呼びます。ロジスティック関数のグラフは，以下のようになることが知られています。

図 1

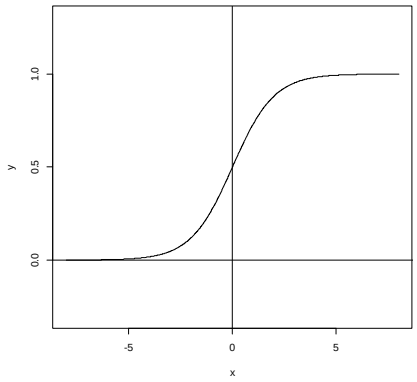


図 1 は、ロジスティック関数が z の単調増加関数になっていることを示しています。すなわち，ロジスティック回帰においては「係数 $\beta>0$ ならば、x の増加は $z=\beta_0+\beta_1x_1+\cdots+\beta_kx_k$ を増やす　⇒　イベント発生確率 p が上昇する」あるいは「正の係数 β は、x 上昇に伴うイベント発生確率の上昇を意味する」と解釈す

ることが出来ます。

ロジスティック回帰とオッズ比の関係：

本節の最後に、ある説明変数 x の値が一単位増加したとき，オッズの値がどう変化するかを考えてみます。いま、$x_2,\ldots,x_k$ の値が一定で，ある説明変数 x_1 の値が一単位増加したとします。x_1 が元の値であったときのイベント発生確率を p 、x_1 の値が一単位増加した後のイベント発生確率を q とすれば、p と q は以下のように書くことが出来ます。

$$p=\frac{\exp[\beta_0+\beta_1x_1+\cdots+\beta_kx_k]}{1+\exp[\beta_0+\beta_1x_1+\cdots+\beta_kx_k]}$$

$$q=\frac{\exp[\beta_0+\beta_1(x_1+1)+\cdots+\beta_kx_k]}{1+\exp[\beta_0+\beta_1(x_1+1)+\cdots+\beta_kx_k]}$$

この式を、 p と q のロジットを使って表すと，以下のように変形出来ます。

$$\log\left(\frac{p}{1-p}\right)=\beta_0+\beta_1x_1+\cdots+\beta_kx_k$$

$$\log\left(\frac{q}{1-q}\right)=\beta_0+\beta_1(x_1+1)+\cdots+\beta_kx_k$$

上の第２式から第１式を引くと，以下のように計算出来ます。

$$\log\left(\frac{q}{1-q}\right)-\log\left(\frac{p}{1-p}\right)=\{\beta_0+\beta_1(x_1+1)+\cdots+\beta_kx_k\}-\{\beta_0+\beta_1x_1+\cdots+\beta_kx_k\}=\beta_1$$

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

$$\Leftrightarrow\frac{q}{1-q}\Big/\frac{p}{1-p}=e^{\beta_1}$$

すなわち，ロジスティック回帰モデルの回帰係数 β を指数変換したとき、e^{β} は説明変数 x が一単位増加した前と後とのオッズ比を示していると言えます。

以上をまとめると，以下のようになります。

- $\beta>0$　　⇒　x の増加はオッズ $\frac{p}{1-p}$ を増やす　⇒　リスクが上昇
- $\beta>0$　　⇒　x の増加はイベント発生確率 p を上昇させる
- $e^{\beta}>1$　　⇒　x の増加の前後のオッズ比 >1　⇒　リスクが上昇

2. ロジスティック回帰におけるデータの要約

Example：Low Infant Birth Weight データ

例として、1986 年にアメリカのマサチューセッツ州スプリングフィールドの病院で生まれた 189 人の幼児のデータを用います。このデータは、出生時体重が 2500g を下回る低出生体重に対する、リスクファクターを探索することを目的としています。以下が変数のリストですが、低出生時体重の発生の有無を示す“low” が被説明変数、そのほかが説明変数となり、$P(\text{low}=1)=p$ を推定するロジスティック回帰モデルを考えます。

low　出生体重が 2.5kg を下回るか否かのダミー変数（0/1).
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+)

Hosmer, D.W. and Lemeshow, S. (1989) *Applied Logistic Regression.* New York: Wiley
Venables, W.N. ana Ripley, B.D. (1999) Modern Applied Statistics with S-PLUS. New York: Springer-Verlag

2.1　連続説明変数の要約

Low Infant Birth Weight データには、二つの連続説明変数 age, lwt が含まれています。age, lwt の数量的要約は以下の通りです。

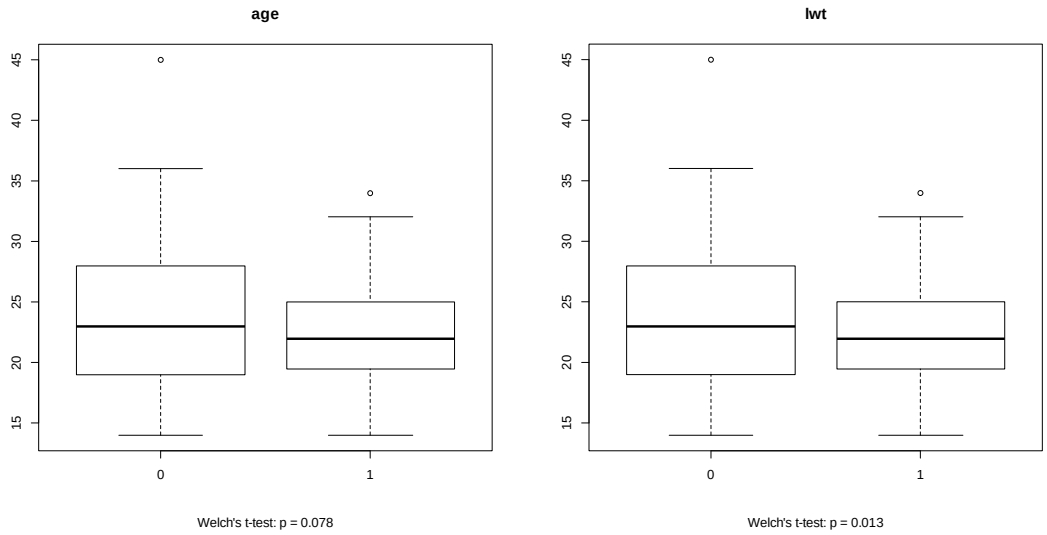
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	SD	IQR
age	14	19	23	23.24	26	45	5.30	7.00
lwt	80	110	121	129.8	140	250	30.58	30.00

連続説明変数の視覚的要約には、被説明変数のイベントの有無によって場合分けしたボックスプロットが有用です。同時に、イベントの有無による二群間の

平均の差を検定するため、**Welch's t-test** を行います。

図 2 によれば、母親の年齢（age）、最終月経期間における母親の体重（lwt）ともに、出生体重 2500g 未満の低出生時体重発生の場合の値が小さくなっていますが、lwt の差が有意（p=0.013）であるのに対して、age の差は有意ではありません（p=0.078）。

図 2



2.2　連続説明変数間の要約

連続変数間の数量的要約は，共分散、相関係数によって行います。また、連続変数間の視覚的要約も，回帰分析の場合と同様、散布図を用いて行います。

図 3

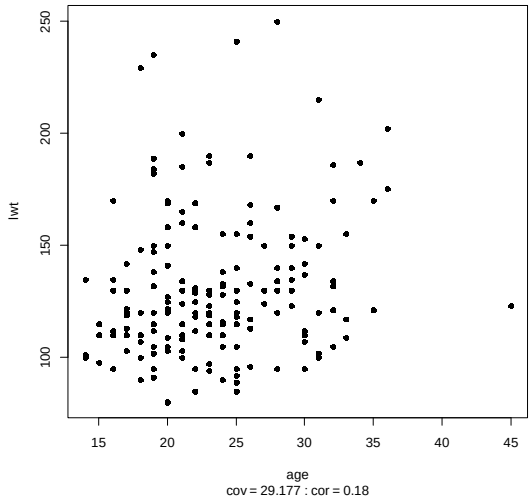


図 3 で示したとおり、共分散=29.177，相関係数=0.18 で弱い正の相関が認められます。回帰分析の場合と同様、ロジスティック回帰の場合も**多重共線性 (multicollinearity)** の発生に注意する必要があります。前回見たとおり，多重共線性とは説明変数の間に強い一時従属（比例）関係がある現象で，推定の結果を不安定にすることが知られています。図 3 の場合、相関係数=0.18 で相関は強いものではありませんので、多重共線性の発生は心配しなくとも良いでしょう。

2.3　離散説明変数の要約

離散説明変数の要約では、被説明変数のイベントの有無と離散説明変数の間の**分割表**と、イベントの有無に関する離散説明変数の水準間の分布の**独立性の検定**を行います。独立性の検定は，可能である限り **Fisher's exact test** によって行います。

```
                race
      white black other
low 0    73    15    42
    1    23    11    25
p-value = 0.079
```

```
               smoke
     0  1
low 0 86 44
    1 29 30
p-value = 0.036
```

```
               ptd
     0  1
low 0 118 12
    1  41 18
p-value = 0
```

```
               ht
     0  1
low 0 125  5
    1  52  7
p-value = 0.052
```

```
               ui
     0  1
low 0 116 14
    1  45 14
p-value = 0.027
```

```
               ftv
     0  1 2+
low 0 64 36 30
    1 36 11 12
p-value = 0.293
```

上記の結果から、喫煙歴有り（smoke = 1)、過去の早産有り（ptd = 1)、子宮炎症有り（ui = 1)、の場合、低出生時体重発生が有意に多く、人種が白人以外、高血圧症有り（ht = 1）の場合もイベント発生が多い傾向が見られました。

3. ロジスティック回帰モデルの推定と検定

以上の予備的なデータのようやくに続き、いよいよロジスティック回帰モデルの未知パラメターの推定と検定に移ります。パラメターの推定は**最尤推定法 (Maximum likelihood estimation)** と呼ばれる方法によって行われますが、推定量の導出の詳細は省略します。

Example：Low Infant Birth Weight データ（続き）

Low Infant Birth Weight データにロジスティック回帰モデルを適用した結果は、例えば以下のように与えられます。

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

ロジスティック回帰分析の結果を検討するときは，以下の点に注意します。

- **回帰係数の推定値**：ロジスティック回帰では、回帰係数の推定値だけでなく推定値の符号にも注意します。P. 5 で見たとおり、$\beta > 0$ ならば x の増加はオッズを増やし、イベントの発生確率とリスクを上昇させます。逆に、$\beta < 0$ ならばイベントの発生確率とリスクを低下させます。上の例では、例えば、age の係数は $\hat{\beta}_1 = -0.037$ で負の値ですから、年齢の上昇は低出生時体重のリスクを低下させます。このことは、第 2.1 節：連続説明変数の要約で age のボックスプロットを描いた際、低出生時体重群で年齢が低い傾向があったことと一致します。
- **回帰係数の有意性検定の p 値**：回帰分析の場合と同様に、ロジスティック回帰分析でも個々の変数の有意性が検定され、検定の p 値が出力されます。上の例では、lwt, race, pptd, ht が有意で有り、smoke も 10%有意水準で有意

でした。

- **回帰係数の信頼区間**：ロジスティック回帰モデルの回帰係数の 95%信頼区間は、以下の公式で与えられる。

  Confidence interval (CI): 回帰係数の推定値±1.96×標準誤差

  たとえば、lwt であれば、$\hat{\beta}_2 = -0.01565, s_{\hat{\beta}_2} = 0.00708$ ですから、

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

  と計算出来ます。
- **オッズ比とオッズ比の信頼区間**：p. 5 で見たとおり、回帰係数を指数変換した e^{β} は説明変数が一単位増加したときのオッズ比に相当します。lwt であれば、

  $$e^{\hat{\beta}_2} = \exp\{\hat{\beta}_2\} = \exp\{-0.01565\} = 0.9845$$

  0.9845 < 1 でオッズ比が 1 未満ですから、lwt（最終月経期間における母親の体重）が増加すると低出生時体重のリスクが低下します。このことは、ボックスプロットで観察した結果とも一致します。オッズ比は回帰係数を指数変換して得られましたから、オッズ比の信頼区間も回帰係数の信頼区間を指数変換して得ることが出来ます。

  $$CI: (\exp\{-0.02953\}, \exp\{-0.00177\}) \Leftrightarrow (0.9709, 0.9982)$$

4. ロジスティック回帰モデルによる予測と判別

前項で推定した回帰係数の推定値を元のロジスティック回帰モデルに代入すれば、イベントの発生確率を予測する予測モデルを得ることが出来ます。

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

そこで、新たな観測値 $x_1, \ldots, x_k$ が得られれば、そのときのイベント発生確率を予測することが出来るようになります。また、イベントの予測確率が得られれば、予測確率が 50%より大きければイベントが発生すると「判別」し、50%未満な

らイベントは起こらないと「判別」することが可能になります。このロジスティック回帰によるイベント発生の予測と判別は、極めて有用な道具になり得ます。

例えば、ある薬剤を投与したとき副作用が発生するか否か、をロジスティック回帰で解析するとします。まず、手元にあるデータにロジスティック回帰モデルを当てはめれば、有意な共変量を探索することで副作用に関与するリスクファクターを同定することが出来ます。さらに、副作用発生確率の予測モデルを作れば、将来新しい患者さんが来たとき、推定したロジスティック回帰モデルの共変量の値を検査して代入することで、薬剤を投与する前に副作用の発生を予測、判別することが出来るようになります。これは、患者さんの個性（＝それぞれの共変量の値）に基づく**個別化医療 (individualized medicine)** につながる研究となるはずです。

5. ロジスティック回帰モデルによる適合度の検定

前々項で検討した回帰係数の推定と検定は、個々の変数の有意性を解析するものでした。しかし、回帰分析において Model utility test を用いてモデル全体の有意性を検討したように、ロジスティック回帰においても

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

というモデル全体の有意性（=適合度）を検定する必要があります。ロジスティック回帰モデルの適合度を検定する方法としてよく用いられるものに、**Hosmer-Lemeshow 検定**と呼ばれるものがあります。

Hosmer-Lemeshow 検定：現在当てはめているモデルが正しいという帰無仮説の元で、以下の手順により検定を行います。

Step1: 前項の方法で、それぞれのサンプルごとにイベントの発生確率を予測する。
Step2: 予測されたイベント発生確率の大小順に、サンプルを事前に定めた数（例えば k=10 個）の群に分割する。
Step3: それぞれの群ごとに、実際に観察されたイベントの発生件数と、予測モ

デルによって判別された「期待」発生件数を比較する。

Step4: $\chi^2 = \sum \frac{(O_i - N_i\hat{\pi}_i)^2}{N_i\hat{\pi}_i(1-\hat{\pi}_i)} \sim \chi^2_{\text{degrees of freedom}=k-2}$

ただし、O_i：第 i 群で観察されたイベントの個数、N_i：第 i 群に含まれるサンプルの個数、$\hat{\pi}_i$：第 i 群に含まれるサンプルの予測確率の平均値、とします。

多くの統計解析ソフトには Hosmer-Lemeshow 検定が実装されていますので、この検定を実行することでロジスティック回帰モデルの適合度を検定します。

Example：Low Infant Birth Weight データ（続き）

統計解析ソフト R version 3.0.2 と R の拡張パッケージ binomTools を用いて、Low Infant Birth Weight データのロジスティック回帰分析を行い、Hosmer-Lemeshow 検定により適合度を検定したところ、検定の p 値は p = 0.6435 となりました。p 値が有意水準 0.05 より大きいので、帰無仮説は棄却されません。現在当てはめているモデルが正しいという帰無仮説が棄却出来なかったことで、モデルの適合度はほぼ良いと結論づけることが出来ます。

6. ロジスティック回帰モデル適用の際の問題点

以上で、ロジスティック回帰モデルの推定方法について議論してきましたが、実際のデータにロジスティック回帰を適用する際にいくつかの困難に直面することがあります。

6.1　多重共線性

これは、連続説明変数間の要約のところでも触れた問題です。説明変数の間に一次従属（＝比例）に近い関係がある場合を多重共線性が発生しているといいますが、その場合、一方の変数の値が決まれば自動的にもう一方の変数の値も決まり「情報の無駄（redundancy)」が存在することになります。数学的にも、多重共線が起こると計算が不安定になり、不合理な推定値が得られたりすることがあります。多重共線を防ぐには、事前の予備的な解析で説明変数相互の間の相関関係を調べ、不必要な変数を外すことが必要になります。

6.2　完全分離

これは､説明変数（あるいは説明変数の組）の値によって､被説明変数のイベントの有無を完全に分離することが出来る場合です。たとえば、ある薬剤を 10mL 以上投与すると必ず副作用が起こるが、投与量が 10mL 未満では副作用は起こらない、といった場合です。このような場合、ロジスティック回帰モデルを当てはめることは出来ません。

完全分離の場合、モデルの推定が出来ないのは数学的な理由によるものですが、考えてみると完全に分離されてしまうということは不確実性が全くないということですから、これは「ロジスティック回帰を使うまでもない状態」と言えます。この場合は､推定や検定は無意味ですから､イベントの有無によって背景因子がどのように違うのか記述統計によって検討することになります。

6.3　外れ値（Outlier）

説明変数の中に､極端に大きいもしくは小さい値がある場合、その他、解析に大きな影響を与えるサンプルがある場合、モデル全体の推定に大きな影響を与える場合があります。そのような外れ値は、データの記録時の誤りや何か例外的な状況で観察される場合があります。そのような外れ値を見つけるには、回帰分析の場合と同様、事前の予備的な解析（特にボックスプロット）と、事後的に得られる残差あるいは標準化残差を用います。

残差（residuals）：$r_i = Y_i - \hat{\pi}_i$．ただし、$Y_i = 0,1; \hat{\pi}_i$：ロジスティック回帰によるイベント発生確率の予測値

標準化残差（standardized residuals）：$r_i \Big/ \sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}$

残差あるいは標準化残差の絶対値が極端に大きいものは、外れ値の可能性があるものとして検討します。ただし、外れ値が見つかったとして、それをどのように処理するかは簡単ではありません。記録時の誤りなどであれば訂正あるいは削除が出来ますが、そうでないときは、外れ値も観察されたデータとして価値のあるものとして扱う必要があります。その際は､外れ値も許容出来るような進んだ統計モデルを用いる必要がありますが､それには専門的な知識が必要で統計の研究者に相談することをおすすめします。

7. 変数選択

回帰分析、ロジスティック回帰分析に限らず広く多変量解析を行う際には、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必要が出てきます。

多数の説明変数の候補から最適な変数の組を選択する**変数選択**では、通常場合の数が膨大になりその全てを探索することは現実的ではありません。実際の探索の過程では、1）変数増加法（forward selection）、2）変数減少法（backward elimination）、3）変数増減法（stepwise procedure）の三通りの方法が用いられます。

さらに、どのような場合に変数を取り込む、あるいは取り除くのか、その方法で変数選択の方法は大きく二つの流儀に分けられます。

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。

まず方法 1 は、以下の手順で行われます。（変数増減法）

Step1-1: まず説明変数を含まない、定数項だけのモデルを考える。そこに i 番目の説明変数 x_i を一つ加えたモデルを考え、x_i の有意性検定の p-value を p_i とする。これを全ての説明変数の候補について行い、p-value の最小値が事前に定めた「投入」の確率より小さければ、最小の p-value に対応する説明変数をモデルに加える。「投入」の確率は $\alpha_1 = 0.1 \sim 0.2$ 程度で、通常の有意水準 0.05 より大きめにとります。

Step1-2: 前の段階で当てはめられたモデルに、さらに説明変数を一つ加えたモデルを考えます。前の Step1-1 と同様に、残った説明変数の候補を加えたときの有意性検定を行い、p-value の最小値が事前に定めた投入確率より小さければ説明変数に加え、投入確率より小さな p-value がなければ forward selection を中断します。新たに投入する説明変数がなくなるまで、このプロセスを繰り返します。

Step1-3: 前段階で投入出来る変数がなくなったところで、今度は既存のモデルに不必要な変数がないか検討します。既存のモデルに含まれた変数一つずつの有意性を検定し、最大の p-value が事前に定めた「除去」の確率より大きかった場合、最大の p-value に対応する変数をモデルから取り除きます。除去確率は $\alpha_2 = 0.1 \sim 0.2$ 程度に設定します。このプロセスを、取り除く変数がなくなるまで続けます。

方法 1 の場合、投入確率、除去確率は解析者が定めるもので、その選択には恣意性が含まれます。

一方、方法 2 ではモデルの当てはまりの良さを測るモデル選択基準を定義する必要があります。よく用いられるモデル選択基準に、以下の二つがあります。

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、L は尤度関数、$-2\log L$ は deviance （逸脱度）と呼ばれるもので回帰分析における残差二乗和に当たるものです。また、p はモデルのパラメターの数、n はサンプル数を表しています。この AIC, BIC のいずれかを、変数増減法で最小化していきます。

上記の変数選択の方法 1, 方法 2 は統計解析ソフトによって使えるものが異なることがあります。また、いずれの方法を選ぶかで変数の選択結果が異なることがありますが、多くの場合、おおむね同じような結果が得られるようです。

変数選択を行う際、よくある質問は以下の二つです。

Q1：方法 1 と方法 2 のいずれを使うべきか？　上に書いた通り、どちらの方法でもあまり変わらない場合が多いので、解析者が自ら選ぶべきです。ただ、強いて言えば、方法 1 が個々の変数の有意性から出発しているのに対し、方法 2 はモデル全体の当てはまりの最適化を目指している点で、方法 2 の方が若干好みかな、という気がします。

Q2：変数選択の結果、注目していた変数が最終的なモデルから外れてしまいま

した。注目したい変数を強制投入しても良いか？　変数選択の方法は、モデルの当てはまり（=被説明変数の予測）の良さを最大にするように設計されています。しかし、多変量解析の目的は予測に限られるものではなく、変数間の関係を推測することも含まれます。その意味で、興味がある変数があれば（最終的にそれが有意になるかは不明だが）モデルに投入して解析することに何の問題もないと思います。

Example：Low Infant Birth Weight データ（続き：変数選択）

Low Infant Birth Weight データに対して、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 を最小化するように変数選択を行いました。

```
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  .
smokeTRUE     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 freedo
```

その結果、母親の年齢（age）と妊娠後 3 ヶ月に医師の診断を受けた回数（ftv）はリスクファクターから脱落し、残りの変数が予測因子として選択されました。

また上の選択結果では、子宮炎症の有無（ui）が残っていますが、個別の有意性検定では p-value = 0.10183 で通常の有意水準 0.05 より大きくなっています。しかし変数選択は、モデル全体（=選択された説明変数の全体）としての予測力を最大にするように変数を選択しますので、ui のような変数もモデルのメンバーとしては意味のある変数として最終的なモデルに含まれるべきものであると言えます。

8. 線型モデルを超えて　ー非線形モデルの世界へー

今回取り上げたロジスティック回帰モデルは、前回検討した線形回帰モデルと同様、モデルが説明変数の一次式に依存するという「線形性の仮定」に基づいて作られています。

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

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

逆に言えば、線形性の仮定という非常に強い仮定を必要とするところに、線形回帰モデル、ロジスティック回帰モデルの理論的な限界がある、ということも出来ます。

ロジスティック回帰では、第１節で見たとおり e^{β} は説明変数 x が一単位増加した前と後とのオッズ比を示していますが、これは x の値にかかわらず x の変化に伴うリスクの変動が一定であることを示しています。

例えば、ある薬剤を投与したときの副作用の有無をロジスティック回帰で解析し、心拍数を説明変数として取り上げたとします。このときロジスティック回帰の線形性を仮定し、e^{β} が心拍数のオッズ比になることを認めたとします。そうすると心拍数にかかわらずオッズ比は e^{β} で一定ですから、心拍数が 50 から 60 に変化しても、100 から 110 に変化しても、150 から 160 に変化しても、副作用発生に対するリスクの変化は変わらない、と仮定したことになります。これは非常に強い、というか非現実的な仮定ではないでしょうか。

回帰モデルにしてもロジスティック回帰にしても、この「線形性の仮定」を認めることで、モデルが簡略になり、様々な理論を展開することが可能になったのです。

しかし、現実に説明変数と被説明変数の関係が非線形であるときは、変数に何らかの変換を施して非線形の方向にモデルを拡張するのは自然なことです。例えば前回取り上げた、回帰モデルの分散安定化と正規性向上のために用いられる Box-Cox 変換は、そのような非線形変換の一つの例になります。

さらに、「線形性の仮定」を緩めモデルに非線形な構造を許したものに、以下の**加法モデル（additive model）、ロジスティック加法モデル（logistic additive model）**があります。

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

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

ここで、 $f_1, \ldots, f_k$ は x の非線形関数で、loess (Locally Weighted Scatterplot Smoother) あるいはスプライン（spline）と呼ばれるクラスの関数になります。 $f_1, \ldots, f_k$ による x の変換は、データに適合するように自動的に行われます。（理論的な詳細は省略します。）加法モデルもロジスティック加法モデルも、より広い**一般化加法モデル（Generalized Additive Model, GAM）**の特別な場合で、統計解析ソフト R には拡張パッケージ gam, mgcv として実装されています。

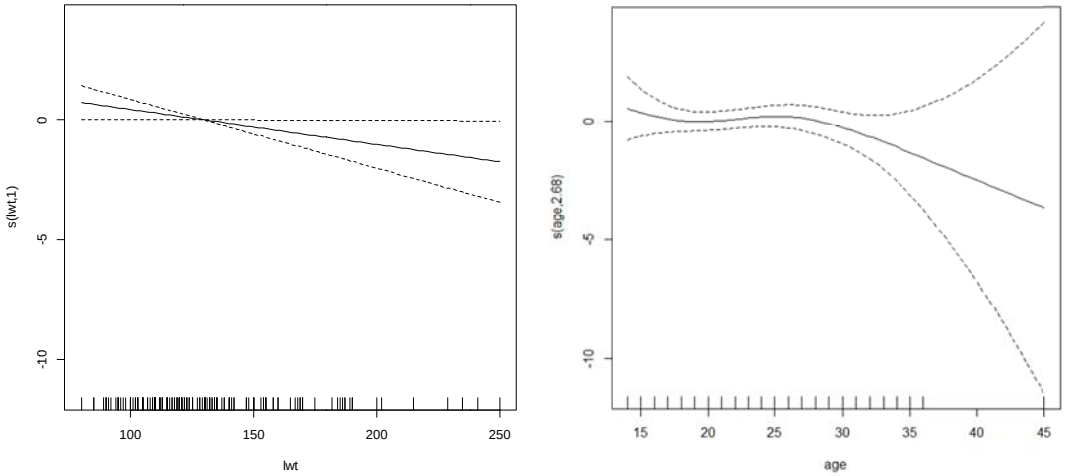
Example：Low Infant Birth Weight データ（続き：ロジスティック加法モデル）

Low Infant Birth Weight データで低出生時体重の発生に対する母親の年齢（age）と最終月経期間における母親の体重（lwt）の非線形な影響を見るため、以下のロジスティック加法モデルを検討した。

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

解析の結果推定された age と lwt の非線形変換を図 4 に示した。解析の結果 lwt に対しては非線形な変換が選択されず線形関係のままであったが、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が見られた（ただし age 配膳として有意ではない）。

図 4



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	*

Take Home Message

1. ロジスティック回帰モデル

ロジスティック回帰モデルの定式化。

- ロジスティック回帰とオッズの関係
- ロジスティック回帰とイベント発生確率の関係
- ロジスティック回帰とオッズ比の関係

2. ロジスティック回帰におけるデータの要約

3. ロジスティック回帰モデルの推定と検定

4. ロジスティック回帰モデルによる予測と判別

5. ロジスティック回帰モデルによる適合度の検定

- Hosmer-Lemeshow 検定

6. ロジスティック回帰モデル適用の際の問題点

- 多重共線性
- 完全分離
- 外れ値

7. 変数選択

- 説明変数に対する逐次的な有意性検定による方法
- モデル選択基準（AIC, BIC）の最適化による方法

8. 線型モデルを超えて

- 線形性の仮定を緩め、モデルに非線形な構造を許容するモデルとして、一般化加法モデルを紹介した。

以上