Logistic回帰モデルにおける変数変換（fractional polynomials ）とモデル適合度診断

古川敏仁
株式会社バイオスタティスティカル　リサーチ　代表取締役

The use of fractional polynominals with Hosmer-Lemeshow Tests and other tools when fitting a logistic regression model

Toshihito Furukawa
President, Biostatistical Research Co. Ltd

要旨

線形モデルにおいては、連続量説明変数の変数形式のモデル適合度診断は、残差プロットや AIC などを用いて、対数変換など適切な変数形式を選択することは比較的容易である。しかし、Logistic 回帰モデルにおいては、応答は[0,1]の 2 値変数であり、推定値と誤差は独立の関係ではない。ゆえに、線形モデルで用いたような手法をそのまま用いることはできない。

そこで、海外では一般的に用いられているが本邦ではあまりなじみのない fractional polynomials という手法を Logistic 回帰モデルの連続量説明変数に適応し、Hosmer-Lemeshow 検定や ROC　AUC などの適合度指標とともにモデル適合度を評価する方法を紹介する。

キーワード：Logistic regression, fractional polynomials, model fitting, Hosmer-LemeshowTest, Numeric Variable

はじめに

一般的に回帰モデルでは、応答 y と説明変数 x の関係は、2 次元上に x-y プロットを描いてみれば一目瞭然であるし、また、残差への系統的なエラーの存在は、x と残差のプロットを作成すればこれまた確認できる。Logistic モデルはリスクの評価や Propensity 解析のような因果推論的解析に多用されるモデルだが、以下の 2 点のような性質から、線形モデルに用いられる fitting の良し悪しを確認するような手法はそのままでは使えない。

1) 応答 y は[0,1]の 2 値変数であり、x-y プロットや残差プロットでは、x の変化に伴う y の変化を確認しにくい
2) 線形モデルでは、ある xj のもとでの条件付期待値 E(Yj|xj)と誤差は独立であるのに対し、Logistic

モデルでは条件付期待値と誤差とには以下の関係が存在する。

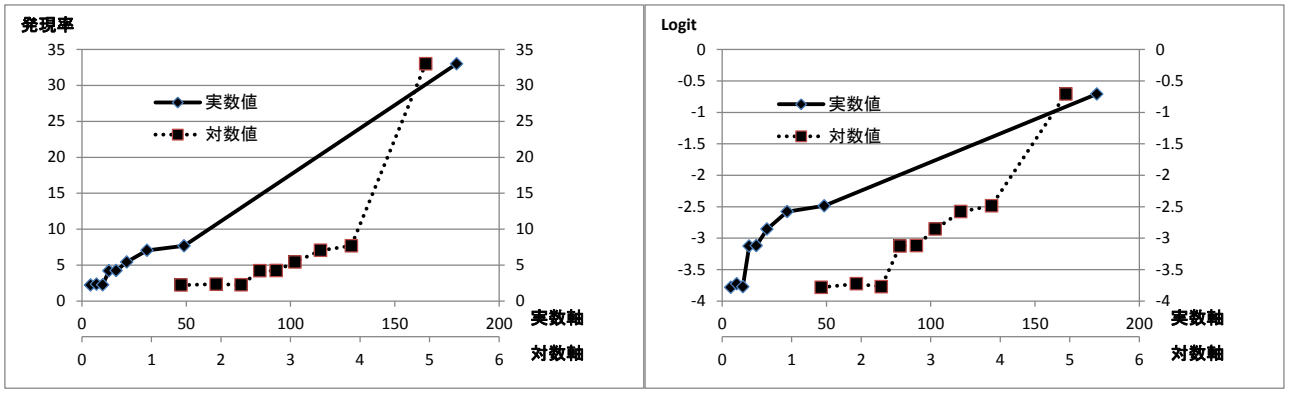
$$\mathrm{Var}(Y_j|x_j) = m_j\pi(x_j)(1-\pi(x_j))$$

ここで mj は x=xj となる例数、 π (xj)は x=xj のときの応答 Y の予測確率

そこで、Logistic モデルでは（xj, yj）の個別データの対応ではなく、x を適切に g 区分した区間での平均値の対応を考えることになる。つまり、応答 Y を Y=1　イベントあり、Y=0　なし　とした場合、x の k 番目の区分の y と x の平均　$\mathrm{mean(y)}_k = \hat{p}_k$、$\mathrm{mean(x)}_k$　y の平均値＝確率の logit 変換値 $\mathrm{logit}(\hat{p}_k) = \log(\frac{\hat{p}_k}{1-\hat{p}_k})$を求めれば、x, y が logistic モデルに適合するならば、$\mathrm{logit}(\hat{p}_k)$と$\mathrm{mean(x)}_k$は直線関係になる。例えば、心血管疾患(CVD)の発生の有無と BNP の関係をグラフ化すると図 1 のようになる。BNP 値は対数正規分布することが知られており、実測値と対数変換値のどちらがモデル適合性があるかが興味の対象となる。図 1　b)は BNP 値を 10 パーセンタイル区分したときの logit（CVD 発症率）と BNP BNP（実測値、対数値）をグラフ表示したものである。実測値、対数値どちらが logit と直線関係に近いのか、あるいはモデル適合性が良いのかグラフからは判別しにくい。そもそも、説明変数の区間区分は任意であり、区分の仕方によって、図表にイメージも違ってくる。そこで、BNP のような連続量説明変数の変数変換形式によるモデル適合性を定量的に評価するような手段が必要となってくる。その手段の 1 つに fractional polynomials や Hosmer-Lemeshow 検定、ROC　AUC などがある。

図 1　心血管疾患(CVD)の発生率と BNP の関係

a)　発症率(%)と BNP（実測値、対数値）　　b)Logit(発症率(%))と BNP（実測値、対数値）



モデル fitting を評価するための統計手法

Fractional Polynomials

fractional polynomials は、連続量説明変数の変数変換形式である。単変量解析、多変量解析どちらの場合でも 1 つの説明変数に適応できる。fractional polynomials は一般的に以下に示す 1 次（J=1)、2 次形式（J=2）の変数変換である。g(x, β)を x の logit 関数とする。

J=1　1 次形式

$g(x, \boldsymbol{\beta}) = \beta_0 + x^p\beta_1$　{p=-2, -1, -0.5, 0, 0.5, 1, 2, 3　ただし、x^0=log(x)}

J=2　2 次形式

$g(x,\boldsymbol{\beta}) = \beta_0 + x^p\beta_1 + x^p\log(x)\beta_2$

$g(x,\boldsymbol{\beta}) = \beta_0 + x^{p1}\beta_1 + x^{p2}\beta_2$　(p1≒p2)

{p,p1,p2=-2, -1, -0.5, 0, 0.5, 1, 2, 3　ただし、x^0=log(x)}

例えば、CVD の例の BNP では、J=1　$g(x,\boldsymbol{\beta}) = \beta_0 + BNP^1\beta_1$ 実測値　p=1、$g(x,\boldsymbol{\beta}) = \beta_0 + \log(BNP)\beta_1$ 対数値　p=0 となる。また J=2 では、p1=0、p2=0 の$g(x,\boldsymbol{\beta}) = \beta_0 + \log(BNP)^1\beta_1 + \log(BNP)^2\beta_2$、p1=0、p2=-0.5 の$g(x,\boldsymbol{\beta}) = \beta_0 + \log(BNP)^1\beta_1 + \frac{1}{\sqrt{BNP}}\beta_2$などが考えられる。

具体的な作業は以下の手順で行う。

① J=1 8 個、J=2 36 個すべてのモデルにデータをあてはめ、-2log(尤度)を計算する。

② J=1 8 個のモデルの-2log(尤度)を小さい順に並べ、最も小さいモデルを Min J=1 モデルとし、その時の-2log(尤度)を Min(-2log(尤度)|J=1)とする。また、変数変換をしない J=1 モデルの-2log(尤度)を-2log(尤度|測定値)とする。

③ 統計量 G(1,p1)= (-2log(尤度|測定値))- Min(-2log(尤度)|J=1)を求め、G(1,p1)を自由度 1 のχ^2検定で検定する*1。すなわち、G(1,p1)>3.84 ならば、Min J=1 モデルの採用を検討し、χ^2検定が有意でなければ変数変換はしない。

④ 続いて J=2 モデルを検討する。J=2 36 個のモデルの-2log(尤度)を小さい順に並べ、最も小さいモデルを Min J=2 モデルとし、その時の-2log(尤度)を Min(-2log(尤度)|J=2)とする。

(1) もし、G(1,p1)が有意ならば、G(p1,(p1,p2))= Min(-2log(尤度)|J=1)-Min(-2log(尤度)|J=2)を求め、G(p1,(p1,p2))を)を自由度 2 のχ^2検定で検定する*1。

(2) もし、G(1,p1)が有意でなければ、G(1,(p1,p2))= -2log(尤度|測定値)-Min(-2log(尤度)|J=2)を求め、G(1,(p1,p2))を)を自由度 3 のχ^2検定で検定する*1。

⑤ J=2 モデルが有意な場合、Max J=2 モデルの採用を検討する。

*1 : G(1,p1)、G(p1,(p1,p2))の分布に関しては、Royston と Altman らのシミュレーション研究(1994)によって、J=1 モデルは実測値モデルに比べて p 乗パラメータ、J=2 モデルは、β2 ならびに p1、p2 パラメータの自由度分の情報量が増加することが報告されている。

Hosmer-Lemeshow 検定

線形モデルでは残差は $(y-\hat{y})$ で示すことができるが、logistic モデルでも以下のような残差を考えることができる。今、説明変数 x、(あるいは複数の説明変数 **x** でも同様) がとり得る値のパターン (covariate pattern) すべてを考え、j 番目の値を xj、xj を持つ例数を mj とし、そのときのイベント例数を yj、モデル確率を$\hat{pj}$とすれば、$\hat{yj} = mj\,\hat{pj}$　となり、残差 r は以下に定義できる。

$$r(yj,\hat{pj}) = \frac{(yj - mj\,\hat{pj})}{\sqrt{mj\,\hat{pj}(1-\hat{pj})}}$$

ここで、$\sqrt{mj\,\hat{pj}(1-\hat{pj})}$は$(yj - mj\,\hat{pj})$の標準誤差であり、$\hat{pj}$と誤差は独立ではないためこのような規準化が必要となる。また、この残差の二乗和は Pearson χ^2 統計量と呼ばれ、自由度 J-(p+1)のカイ二乗分布に従

う。（J は covariate pattern 数、p は変数の数）

$$X^2 = \sum_{j=1}^{J}\left(r(yj, \hat{pj})\right)^2$$

この統計量も、モデルへの当てはめの良さを示す 1 つの指標にはなるが、連続量変数では 1 つの covariate pattern に属する例数が 1 に近くなるため、カイ二乗分布への近似に問題が生じることになる。そこで、x を適切に g 区分した区間（一般的には g=10、x 値の 10 パーセンタイル区分が用いられる。）で以下の統計量$\hat{C}$を求める。

$$\hat{C} = \sum_{k=1}^{g} \frac{(o_k - n_{k'}\overline{p_k})^2}{n_{k'}\overline{p_k}(1-\overline{p_k})}, \quad o_k = \sum_{j=1}^{ck} y_j, \quad \overline{p_k} = \sum_{j=1}^{ck} \frac{mj\,\widehat{pj}}{n_{k'}}$$

ck：k 区分の covariate pattern 数

mj：k 区分の j 番目の covariate pattern に属するデータ数

n_k'： k 区分のデータ数

$\hat{C}$は Hosmer と Lemeshow のシミュレーション研究から自由度 g-2 のカイ二乗分布へ近似できることが報告されており、χ^2（g-2）に基づく検定は Hosmer-Lemeshow 検定と呼ばれ、$\hat{C}$のχ^2（g-2）近似特性は良いことが知られている。

SAS では、モデルオプションに LACKFIT オプションを指定すれば Hosmer-Lemeshow 検定結果が、SCALE=p と AGGREGATE オプションを指定すれば Pearson χ^2 検定結果が出力される。SAS の Hosmer-Lemeshow 検定の出力は、10 パーセンタイルごとに区分されたグループ別の、例数、イベント観察例数、期待値が出力される（表 1）。そのため表 1 を見れば、モデル適合性の良い区間と問題がある区間が一目で確認できるようになっている。

表 1　CVD データにおける対数変換値の Hosmer と Lemeshow 検定の分割

グループ	全体	CVD = 1		CVD = 0	
		観測値	期待値	観測値	期待値
1	312	209	220.5	103	91.5
2	324	298	285.2	26	38.8
3	301	281	276.7	20	24.3
4	294	278	276.8	16	17.2
5	283	271	270.1	12	12.9
6	285	273	274.6	12	10.4
7	311	304	302.2	7	8.8
8	340	332	333.2	8	6.8
9	124	120	122.2	4	1.8
10	549	538	542.5	11	6.5

Hosmer と Lemeshow の適合度検定

カイ 2 乗	自由度	Pr > ChiSq
14.5912	8	0.0676

ROC AUC

ROC AUC は、検査 x に関しすべての閾値をカットオフとし、感度（イベント例中正しくイベントと判定された割合）を縦軸に、1―特異度（非イベント例中正しく非イベントと判定された割合）を横時にとったときに描かれる ROC 曲線の曲線下面積である。AUC の持つ意味は以下の 2 つである。

1) 感度、特異度の期待値、すなわち検査の診断能を示す数値であり、AUC=0.5 の場合まったく診断能のない検査であることを示す。一般的に 0.7 以上であれば診断能がある、0.8 以上で良い診断能、0.9 以上で際立って良い診断能といわれる。
2) イベントと非イベントのすべての組み合わせを考えた場合、イベントに属すると予測するモデル確率がイベントの方が高い組み合わせの割合。

ROC AUC はノンパラメトリック統計量であるため､J=1 のような変数変換ではその値は変わらない｡また、J=2 の場合において、モデル適合性は増すが、covariate pattern 数の減少のため AUC は減少する場合もある。AUC はモデルの診断能の評価指標ではあるがモデル適合性の評価指標ではない。しかし、多変量解析において、交互作用や変数の組み合わせを評価する場合、モデルの診断能とモデル適合性が同一方向となる場合があり、モデル適合性の重要な指標となる場合がある。

SAS logistic プロシジャでは「予測確率と観測データの応答との関連性」出力の c 統計量が AUC に該当する。

SAS コーディング例と結果の解釈

SAS コーディング

作成した SAS マクロ（添付Ⅰ）では、fractional polynomials の J=1、J=2 のすべての変数変換形式に対して、-2log(尤度)、Hosmer-Lemeshow 検定の$\hat{C}$、自由度、p 値、Pearson χ^2検定のχ^2値、自由度、p 値、ROC AUC を算出し、J=1、J=2 グループ別に-2log(尤度)の小さい順にソートしている。

出力の利用について

CVD データに関して、表 2 に J=1 グループの出力結果が、表 3 に J=2 のグループの-2log(尤度)が小さい 4 つモデルの出力結果を示している。

表 2　CVD データにおける fractional polynomials 出力　J=1

		Hosmer-Lemeshow検定			Pearson χ^2検定			ROC
p(次数)	-2LogLR	χ 2	DF	p値	χ 2	DF	p値	AUC
log	1319.8	14.59	8	0.0676	0.99	164	0.5055	0.774
0.5	1324.7	5.46	8	0.7073	1.30	164	0.0063	0.774
−0.5	1376.4	69.24	8	<.0001	1.45	164	0.0001	0.774
1	1379.7	27.54	8	0.0006	3.11	164	<.0001	0.773
−1	1440.2	133.76	8	<.0001	2.23	164	<.0001	0.774
2	1485.2	119.12	8	<.0001	5.44	164	<.0001	0.732
−2	1506.4	220.15	8	<.0001	3.40	164	<.0001	0.769
3	1525.6	197.54	6	<.0001	4.15	164	<.0001	0.700

表 3　CVD データにおける fractional polynomials 出力　J=2

		Hosmer-Lemeshow検定			Pearson χ^2検定			ROC
p(次数)	-2LogLR	χ 2	DF	p値	χ 2	DF	p値	AUC
log,-0.5	1310.9	4.41	8	0.8186	1.00	163	0.4949	0.774
log,-1	1311.1	4.89	8	0.7689	0.98	163	0.5448	0.772
log,0.5	1311.7	4.96	8	0.7618	0.99	163	0.5069	0.774
log,log*log	1312.1	4.58	8	0.8014	1.02	163	0.4227	0.774

J=1 グループにおいては、対数変換値が最も小さな-2log(尤度) 1319.8 であり、p=1 すなわち実測値の BNP 値の 1379.7 との差は 59.9>3.84 であり、1 次変換では BNP 値は対数変換がモデル適合性が有意に高いことが示されている。Hosmer-Lemeshow 検定の p 値は 0.0676、Pearson χ^2 検定の p 値は 0.5055 であり、Pearson χ^2 検定では、モデル適合に問題となるような系統的なエラーは検出されないが、Hosmer-Lemeshow 検定では有意ではないが p 値は小さい。これは図 1　b)を見れば明らかなように、BNP 値は 20 までほとんど CVD リスクは変化せず、20 以降、対数 BNP と直線的に logit が増加しているため、その分の系統的エラーが検出されたためだと思われる。また、Hosmer-Lemeshow 検定の出力（表 1）からも、BNP≦20 に該当するグループ 1,2,3 のイベント数とモデル期待値の乖離が大きいこともこれを示唆している。つまり、Pearson χ^2 検定よりも Hosmer-Lemeshow 検定の方が、変数形式のモデル適合度異常に関する感度が高いことを示唆している（表 2）。しかし、2 行目　平方根変換においては、-2log(尤度) 1324.7 は対数変換値より有意に悪いモデル適合性だが、Hosmer-Lemeshow 検定は P=07073 と有意ではなくなり、Pearson χ^2 検定 P=0.0063 とこちらは、有意となっている。これは、Hosmer-Lemeshow 検定に代表されるモデル適合度検定はどのシチュエーションでも均一な性能を示すものではなく、モデル適合性は複数のツールを用いて総合的に判断するものであることを示している。J=2 の組み合わせでは(p1,p2)=(0,-0.5)の組み合わせが最も大きな-2log(尤度)をもち、J=1 の最大値の対数モデルとの差は 1319.8-1310.9=8.9 であり、自由度 2 の α =0.05 のカイ二乗分布点 5.99 よりも大きな値なので、J=2 の 2 次モデルが BNP の CVD リスク予測には有用であることが示唆されている（表 3)。なお、(p1,p2)=(0,-0.5)においては、Hosmer-Lemeshow 検定 P=05146、Pearson χ^2 検定 P=0.6067 とも有意ではなく、2 次モデルの当てはめのよいことが示されている。

さて、2 次モデルが選択される理由は先に示した BNP 値は 20 までほとんど CVD リスクは変化せず、20 以降、対数 BNP と直線的に logit が増加しているためであると想像され（図 1　b）る。また、この結果は、BNP<=20 のデータを除いた fractional polynomials の結果で 2 次形式が有意にならないことからも確認されている。
しかし、$\text{logit} = \beta 0 + \log(BNP)\,\beta 1 + \frac{1}{\sqrt{BNP}}\beta 2$というモデルは臨床的に理解しづらいため、実際の CVD 予測モデルでは$\text{logit} = \beta 0 + \log(BNP|BNP > 20)\,\beta 1 + I(BNP \le 20)\beta 2$という、Index 関数を用いたモデルを採用した。

fractional polynomials　おわりに

- fractional polynomials は, logistic モデルの連続量変数の変数形式の選択において非常に強力なツールであり、今回は、その原理、SAS プログラム見本、結果の解釈の仕方を紹介した。
- 今回の事例は単変量解析の事例であるが、多変量解析においても同様に fractional polynomials は強力なツールであり、特に交互作用や変数の組み合わせに関して ROC AUC が重要な役割を持つようになる。

- -2log(尤度)、Hosmer-Lemeshow 検定、Pearson χ^2検定は、連続量変数の変数変換の適合度診断にそれぞれ強力なツールとなるが、また、それぞれ限界もあり、グラフ表示と合わせ総合的に変数変換の適合度診断を行う必要がある。
- 連続量変数の変数変換の形式は最終的には臨床的観点から決定する必要がある。
- fractional polynomials のような変数変換選択方法はデータ依存性があり、external validation でその妥当性を確認する必要がある。

添付 1：SAS コーディング

```
/*   fractional polynomials 変数の作成 */
%MACRO DDD;     %DO I=1 %TO 8; J&I.&I.=J&I.*J5;     %END;
%MEND;

%LET VAR= fractional polynomials 変数名 P;
DATA ANL;   SET SAS データセット;
 J1=&VAR*&VAR*&VAR;     J2=&VAR*&VAR;    J3=&VAR;    J4=SQRT(&VAR);
 J5=LOG(&VAR);          J6=1/J4;         J7=1/J3;    J8=1/J2;
 %DDD;  RUN;

/* 統計量出力マクロ */
%MACRO LGST;
 PROC DATASETS;DELETE FIT AS GD HS;RUN;
QUIT;
ODS OUTPUT Association=AS FitStatistics=FIT LackFitChiSq=HS  ;
proc logistic data=anl;
 model CVD(event='1')=&var/ LACKFIT ;  ;
run;   QUIT;
ODS OUTPUT   GoodnessOfFit=GD;
proc logistic data=anl;
 model CVD(event='1')=&var/SCALE=p AGGREGATE;  ;
run;       QUIT;

DATA FIT; SET FIT;  IF Criterion='-2 Log L'   ; RUN;
DATA AS; SET AS; IF Label1='組'; RUN;
DATA GD; SET GD;   IF Criterion='Pearson';  RENAME DF=GDDF  ProbChiSq =GDProbChiSq  ; RUN;
DATA DAT&N ;  MERGE  FIT AS GD HS; N="&N";   RUN;
DATA  &dat; SET &dat DAT&N; RUN;
%MEND LGST;
```

```
/* ルーチン*/
PROC DATASETS;DELETE FP FPB FPC;RUN;
DATA FP;CC=1;RUN;
DATA FPB;CC=1;RUN;
DATA FPC;CC=1;RUN;
 %MACRO FP;
  %DO N=1 %TO 8;
  %let var=&j&n;%let dat=fp;%lgst;
  %let var= J&N J&N.&N;%let dat=fpb;%lgst;
  %LET S=%EVAL(&N+1);
  %DO NN=&S %TO 8;
%let var= J&N J&NN;%let dat=fpc;%lgst;
  %END; %END;
 %MEND FP;
%FP;
/*出力  */
PROC SORT DATA=FP;
 BY DESCENDING InterceptAndCovariates ;RUN;
filename ex dde "excel|シート名 1!R6C2:R280C10";
  data _null_;   SET FP;   file ex ;
    put N  InterceptAndCovariates  ChiSq  DF  ProbChiSq
      ChiSqDivDF GDDF GDProbChiSq nValue2 ;
  run;

DATA FP2;  SET FPB FPC;    RUN;
PROC SORT DATA=FP2;   BY DESCENDING InterceptAndCovariates ;RUN;
filename ex dde "excel|J2!R6C2:R42C10";
  data _null_;     SET FP2;  file ex ;
    put N  InterceptAndCovariates  ChiSq  DF  ProbChiSq
      ChiSqDivDF GDDF GDProbChiSq nValue2 ;
  run;
```

参考文献

1) Applied Logistic Regression Second Edition, David W. Hosmer, Stanley Lemeshow, Wiley 2000
2) Royston, P., and Altman, D. G. (1994) Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling (with discussion). Applied Statics, 43, 429-467
3) Hosmer, D. w. and Lemeshow, S.(1980). A goodness-of-fit test for the multiple logstic regression model. Communication in Statics, A10, 1043-1069