复杂曲线回归分析及其应用

谷恒明[1]，胡良平[1,2*]

（1. 军事医学科学院生物医学统计学咨询中心，北京 100850；
2. 世界中医药学会联合会临床科研统计学专业委员会，北京 100029
*通信作者：胡良平，E-mail：lphu812@sina.com）

【摘要】 本文目的是介绍曲线方程的两种拟合方法（即直线化法与直接拟合法）之间的区别。通过对一个实例中散点图为“S 型曲线”的资料采用直线化法与直接拟合法拟合 Logistic 曲线回归方程，对所得到的拟合结果进行比较，发现基于粗估值再采用 SAS 中的 NLIN 过程直接拟合 Logistic 曲线回归方程，可以得到更精确的拟合效果。

【关键词】 非线性回归分析；Logistic 曲线回归方程；迭代计算；相关指数
中图分类号：R195.1 文献标识码：A doi：10.11886/j.issn.1007-3256.2017.06.004

Complex curve regression analysis and its application

Gu Hengming[1], *Hu Liangping*[1,2*]

(1. *Consulting Center of Biomedical Statistics, Academy of Military Medical Sciences, Beijing* 100850, *China*;
2. *Specialty Committee of Clinical Scientific Research Statistics of World Federation of Chinese Medicine Societies, Beijing* 100029, *China*
** Corresponding author: Hu Liangping, E-mail: lphu812@sina.com*)

【Abstract】 The purpose of this paper is to introduce the difference between the two fitting methods (linearization and direct fitting) of curve equation. By fitting the Logistic curve regression equation with the linearized method and the direct fitting method for an example in which the scatter plot is "S-shaped curve". Compared the fitting results producing from the two methods mentioned above, it was found that the more accurate fitting result can be gotten by means of providing the crude estimators of the parameters in the regression equation and then directly fitting Logistic curve regression equation through the NLIN procedure in SAS/STAT.

【Keywords】 Nonlinear regression analysis; Logistic curve regression equation; Iteration calculation; Correlation index

1 概 述

1.1 问题的提出

本刊本期的前一篇文章《简单曲线回归分析及其应用》介绍了用曲线直线化的方法实现几个常用初等函数曲线的曲线回归或曲线拟合的方法。然而，人们可能会提出下面两个问题。问题一：曲线直线化方法与直接进行曲线拟合的效果一样吗？问题二：是否所有的函数曲线都可以通过曲线直线化构建曲线回归方程？

对第一个问题的回答是否定的。其理由如下：曲线直线化的原理是对原因变量或/和结果变量进行变换，使变换后的变量呈直线关系，然后，以最小二乘法来拟合变换后的原因变量与结果变量之间的直线关系。这样所得的结果是使变换后的结果变量与直线回归方程计算所得的结果变量的预测值（即变换后的结果变量的预测值）之间的离差平方和最小，却并非原结果变量与其预测值之间的离差平方和最小。以 Logistic 曲线为例，原因变量与结果变量的关系式为 $y=L+\frac{K}{1+ae^{bx}}$，现对结果变量 y 进行变量转换，令 $y'=\ln\frac{K-(y-L)}{y-L}$，则 Logistic 曲线回归方程可转化为 $y'=bx+\ln a$。设 y' 的预测值为 $\hat{y}'$，曲线直线化通过最小二乘法求得的曲线回归方程仅确保 $\sum(y'-\hat{y}')^2$ 值最小，而非 $\sum(y-\hat{y})^2$ 最小。因此，其拟合精度较直接拟合曲线所对应的精度要差一些。

对第二个问题的回答也是否定的。其理由如下：事实上，能够通过直线化来实现曲线拟合的曲线类型是很少的，绝大多数函数曲线是无法通过简单直线化法来实现曲线拟合的。例如下面的两个函数曲线就无法通过直线化法进行曲线拟合。

二项型指数函数是由两个指数函数项相加而构成的函数表达式。此函数表达式所描绘出的曲线称为二项型指数曲线，见式（1）。

$$y=A\times e^{\alpha t}+B\times e^{-\beta t} \quad (1)$$

又例如：三项型指数函数是由三个指数函数项相加而构成的函数表达式。此函数表达式所描绘出的曲线称为三项型指数曲线。此曲线常用于研究三室模型药物静脉注射或二室模型药物血管外给药后

项目基金：国家高技术研究发展计划课题资助（2015AA020102）

血药浓度与时间的关系，见式(2)。

$$y = Ne^{-k_x} + Le^{-\alpha x} + Me^{-\beta x} \quad (2)$$

1.2 解决方案

SAS 软件提供了 NLIN 和 NLMIXED 过程，可用于非线性曲线的拟合。但在拟合时需要读者给出未知参数的初始值。若初始值偏离真实值较远，则可能拟合过程无法收敛或拟合曲线并非最优。所以，较为合适的方法是通过曲线直线化求出参数的粗估值，然后将其作为初始值引入 NLIN 或 NLMIXED 过程进行迭代计算，以寻找拟合效果更好的曲线回归方程。然而，对于类似上面的式(1)和(2)那样无法实施直线化的函数方程式，估计参数的粗估值可能就带有很大的盲目性了。具体如何实现，请参见相关参考文献[1-2]。

由于直接进行曲线拟合问题难度较大，再考虑到文章篇幅有限，故本文仅介绍一个最常用的曲线拟合的实例。

2 实例解析

2.1 问题与数据

【例 1】某研究者欲分析某县疟疾发病的季节性特点，观测了某县 1961 年 - 1996 年疟疾的月累计发病率(1/10 万)，结果见表 1。试对该资料进行曲线拟合。

表 1 某县疟疾月累计发病率

月份(月)	累计发病率(1/10 万)
1	0.555
2	1.295
3	2.751
4	11.116
5	24.879
6	43.476
7	69.297
8	97.037
9	114.631
10	121.645
11	124.412
12	125.619

2.2 如何用 SAS 实现计算

【分析与解答】图 1 散点图显示，资料的趋势近似 S 型曲线，故可采用 Logistic 曲线方程拟合此资料。曲线上限为 130，下限为 0，上下渐近线之间的垂直间隔为 130，SAS 程序如下：

```
%let ul = 130;                    /*1*/
dataabc;                          /*2*/
  do x = 1 to 12;
  input y@@;
  z = log((&ul - y)/y);
  output;
  end;
  cards;
  0.555  1.295  2.751  11.116
  24.879 43.476 69.297 97.037
  114.631121.645124.412125.619
  ;
run;
ods html;
proc gplot;                       /*3*/
  plot y*x/haxis = 0 to 12 vaxis = 0 to &ul by 10;
  symbol value = dot;
run;
proc reg data = abc outest = est(keep = Intercept x);
  model z = x;                    /*4*/
run;quit;
data set1;                        /*5*/
  set est;
  call symput('a1',exp(Intercept));
  call symput('b1',x);
run;
proc nlin data = abc;             /*6*/
  parms K = &ul a = &a1 b = &b1;
  model y = K/(1 + a*exp(b*x));
  output out = set2 p = yp r = resid;
run;
proc sql;                         /*7*/
  create table set3 (sum num,css
num);
  insert into set3 select
sum(resid**2),css(y) from set2;
quit;
data set4;                        /*8*/
  set set3;
  r2 = 1 - sum/css;
run;
proc print data = set2 round;     /*9*/
```

```
run;
proc print data = set4;      /*10*/
run;
ods html close;
```

【程序说明】程序中第 1 步是设立宏变量 ul，表示曲线的上、下渐近线的垂直间距，其值为 130，以便于后面调用。第 2 步是创建数据集，x 表示月份，y 表示疟疾月累计发病率，z 是对 y 进行 logit 变换所得的变量，以便后面进行曲线直线化分析。第 3 步是绘制散点图，以发现散点分布趋势，本资料散点趋势近似 S 型曲线。第 4 步对变量 z 和 x 进行直线回归分析，将所得直线的截距和斜率保存在数据集 est 中。第 5 步对数据集 est 进行操作，将以自然对数为底、以截距为指数的数值赋给宏变量 a1，将斜率赋给宏变量 b1。当然，也可直接将截距赋给宏变量 a1，将斜率赋给宏变量 b1，但第 6 步中 model 语句给出的曲线方程需改为“y = K/(1 + exp(a + b * x))”。第 6 步是调用 NLIN 过程来拟合 Logistic 曲线，宏变量 ul、a1、b1 的值分别赋给变量 K、a、b 作为初值，并将所得曲线对各观测拟合的预测值及残差输出到数据集 set2 中，预测值以变量 yp 表示，残差以 resid 表示。第 7 步将数据集 set2 的内容输出到output窗口。第7步是调用sql过程，建立一个表set3，包括 sum 和 css 两个变量，然后将数据集 set2 中所有观测 resid 的平方和赋给 set3 中的 sum 变量，将 set2 中所有观测 y 变量的离均差平方和赋给 css。第 8 步是对数据集 set3 进行操作，计算新的变量 r2，即相关指数。第 9 步是将数据集 set2 的内容输出到output窗口，round 选项用来指定数据最多保留两位小数。第 10 步是将数据集 set4 的内容输出到output窗口。

2.3　SAS 输出结果及其解释

【主要输出结果及解释】

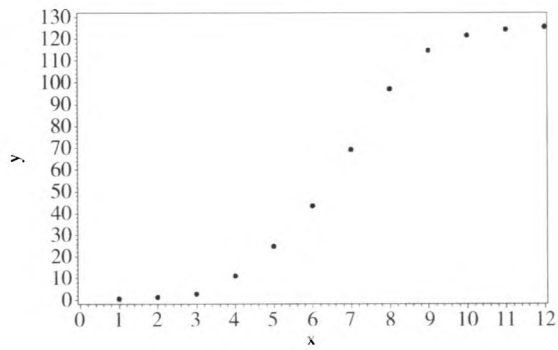

图 1　y 随 x 变化的散点图

图 1 是 SAS 程序绘制的散点图。可以看出，散点近似 S 型曲线趋势，上限为 130，下限为 0。

Analysis of Variance Dependent Variable: z					
Source	DF	SumofSquares	MeanSquare	F Value	Pr > F
Model	1	102.44535	102.44535	658.11	<.0001
Error	10	1.55666	0.15567		
Corrected Total	11	104.00201			

Root MSE	0.39455	R - Square	0.9850
Dependent Mean	0.50213	Adj R - Sq	0.9835
Coeff Var	78.57356		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	6.00377	0.24283	24.72	<.0001
x	1	-0.84640	0.03299	-25.65	<.0001

这是曲线直线化分析的结果，即变量 z 与 x 之间直线回归分析的结果。截距为 6.00377，斜率为 -0.8464，它们与 0 之间的差异均有统计学意义，检验统计量 t 值分别为 24.72、-25.65，P 均 <0.0001。直线回归方程为 $z = 6.00377 - 0.8464x$，转换成 y 与 x 的关系，即为：$y = 130/(1 + e^{6.00377 - 0.8464x})$。此时，残差平方和为 194.02219，相关指数 $R^2 = 0.99740$。

The NLIN Procedure

Dependent Variable y Method: Gauss - Newton

Iterative Phase				
Iter	K	a	b	Sum of Squares
0	130.0	405.0	-0.8464	249.8
1	127.3	374.8	-0.8809	21.0766
2	127.8	394.3	-0.8881	19.3877
3	127.8	394.5	-0.8879	19.3807
4	127.8	394.6	-0.8879	19.3807
5	127.8	394.6	-0.8879	19.3807

NOTE: Convergence criterion met

Estimation Summary	
Method	Gauss - Newton
Iterations	5
R	2.53E-6
PPC(a)	1.044E-6
RPC(a)	0.000015
Object	1.416E-9
Objective	19.38066
Observations Read	12
Observations Used	12
Observations Missing	0

Note: An intercept was not specified for this model

这是 NLIN 过程拟合曲线模型的有关信息。迭代方法为 Gauss - Newton 法，以 $K=130, a=405, b=-0.8464$ 为初始值，迭代了 5 次，得到了收敛的结果。

Source	DF	Sum of Squares	Mean Square	F Value	Approx Pr > F
Model	3	76037.4	25345.8	11770.1	<.0001
Error	9	19.3807	2.1534		
Uncorrected Total	12	76056.8			

Parameter	Estimate	Approx Std Error	Approximate 95% Confidence Limits	
K	127.8	1.0737	125.4	130.2
a	394.6	65.1746	247.1	542.0
b	-0.8879	0.0263	-0.9474	-0.8284

Approximate Correlation Matrix			
	K	a	b
K	1.0000000	-0.5086692	0.6148707
a	-0.5086692	1.0000000	-0.9798514
b	0.6148707	-0.9798514	1.0000000

这是 NLIN 过程拟合模型的结果。对模型的检验结果为 $F=11770.1, P<0.0001$，说明所拟合的模型是有统计学意义的。模型的三个参数取值分别为：$K=127.8, a=394.6, b=-0.8879$。所以，本资料所拟合的 Logistic 曲线回归方程为：

$\hat{y}=127.8/(1+394.6\times e^{-0.8879x})$。

Obs	x	y	z	yp	resid
1	1	0.56	5.45	0.78	-0.23
2	2	1.30	4.60	1.88	-0.59
3	3	2.75	3.83	4.49	-1.73
4	4	11.12	2.37	10.38	0.74
5	5	24.88	1.44	22.60	2.28
6	6	43.48	0.69	43.83	-0.35
7	7	69.30	-0.13	71.46	-2.17
8	8	97.04	-1.08	96.50	0.53
9	9	114.63	-2.01	112.76	1.87
10	10	121.65	-2.68	121.16	0.49
11	11	124.41	-3.10	124.99	-0.58
12	12	140.62	-3.36	126.64	-1.02

Obs	sum	css	r2
1	19.3807	30827.91	0.99937

这是曲线拟合效果的结果。首先给出了曲线回归方程对各观测的拟合结果，yp 列为曲线对各观测的预测值，resid 为预测值与观测值之间的残差。然后给出了对拟合整体效果的评价情况，残差平方和为 19.3807，相关指数 $R^2=0.99937$。

由此可知，在粗估值基础上直接进行 Logistic 曲线拟合，可获得更好的拟合效果。

参考文献

[1] 胡良平，高辉. 非线性回归分析与 SAS 智能化实现[M]. 北京：电子工业出版社，2013：68-119.

[2] 胡良平. 科研设计与统计分析[M]. 北京：军事医学科学出版社，2012：399-426.

（收稿日期：2017-12-03）

（本文编辑：陈 霞）