维普资讯 http://www.cqvip.com

文章编号:0253－3626(2003)06－0754－04

非线性生物模型回归参数计算的一个新方法及应用

张世强

(重庆医科大学基础医学院数学教研室,重庆 400016)

【摘 要】目的:研究变量代换对非线性生物模型回归参数的干扰,寻找消除或减少该干扰的新方法。方法:利用空间变换、泰勒(Taylor)级数和加权处理的组合方式消除或减少变量代换带来的严重影响。结果:给出一个计算非线性生物模型回归参数的新方法。结论:新方法消除或减少了变量代换带来的严重影响,保留了变量代换便于计算的优点;不但显著提高了非线性生物模型回归参数的精度,而且还能发现原始数据中隐藏的问题。

【关键词】非线性模型;最小二乘法;空间变换;非线性参数估计

【中国图书分类法分类号】O212 【文献标识码】A 【收稿日期】2003－01－17

A new method for calculating regression parameters of nonlinear biology model and its application

ZHANG Shiqiang

(*Department of Mathematics, College of Basic Medical Sciences, Chongqing Medical University*)

【Abstract】 *Objective*: To study the disturbance creatd by the variable transformation within regression parameters of nonlinear biology models and find a new method removing or reducing the disturbance. *Methods*: To use the space transformation, Taylor series and weight method to remove or reduce the serious disturbance created by the variable transformation within regression parameters of nonlinear biology models. *Results*: Giving a new method calculating regression parameters of nonlinear biology models. *Conclusion*: The new method can remove or reduce the serious disturbance created by the variable transformation and remain the advantage of variable transformation. It can not only improve regression precision of the nonlinear biology models but also can find problems hidden in the original data.

【Key words】 Nonlinear model; Least square; Space transformation; Estimation of nonlinear parameter

在生命科学中,科学实验所获得的数学模型绝大多数是非线性的。由于非线性回归模型比线性回归模型复杂,不太容易计算出回归参数,这极大地阻碍了非线性回归分析在生命学科及其它学科中的应用。虽然已有多种形式的非线性回归参数计算的方法,如直接搜索法,梯度法,截尾牛顿法,Hooke－Jeeves 法,变尺度法以及 Nelder－Mead 法等计算方法。但这些算法所要求的条件较苛刻,专业针对性强。故目前国内外大多数应用回归分析的教科书,均采用经典的 Gauss－Newton 法。

非线性回归模型的一般形式为:

$$\left.\begin{aligned} y &= f(x_1, x_2, \cdots x_m, \alpha_1, \alpha_2, \cdots \alpha_l) + e \\ e &\sim N(0, \sigma^2) \end{aligned}\right\} \quad (1)$$

其中自变量 $x=(x_1, x_2, \cdots x_m) \in R^m$,参数 $\alpha=(\alpha_1, \alpha_2, \cdots \alpha_l) \in R^l$,因变量 $y \in R^1$。含参数的多元函数 $f(x_1, x_2, \cdots x_m; \alpha_1, \alpha_2, \cdots \alpha_l)$ 是关于自变量 $x=(x_1, x_2, \cdots x_m) \in R^m$ 的非线性函数。非线性回归分析首先要解决的问题是如何求得参数。即已知实验数据集合:

$$\{((x_1, x_2, \cdots x_m)_i, y_i) \mid i=1,2,\cdots,n)\},$$

如何用最小二乘法求出参数 α 的最佳估计值 $\hat{\alpha}=(\alpha_1, \alpha_2, \cdots \alpha_l) \in R^l$,满足:

$$S(\hat{\alpha}) = \min_{\alpha} S(\alpha) = \min_{\alpha} S(\alpha_1, \alpha_2, \cdots \alpha_l),$$

其中 $S(\alpha) = \sum_{i=1}^{n} (y_i - f(x_1, x_2, \cdots x_m; \alpha_1, \alpha_2, \cdots \alpha_l))^2$。

国内外大多数应用回归分析的教科书所采用 Gauss－Newton 法的计算过程是:首先取 $f(x_1, x_2, \cdots x_m, \alpha_1, \alpha_2, \cdots \alpha_l)$ 关于参数 α 的 Taylor 级数的常数项和线性项,然后利用线性最小二乘法,再通过迭代逐次逼近参数 α 解。然而对于非数学专业的生物工

基金项目:重庆医科大学科技基金资助课题(XB200221)

作者介绍:张世强(1958－),男,副教授,
主要研究方向:医药学数据处理与建模。

维普资讯 http://www.cqvip.com

作者来说，Gauss－Newton 法计算公式复杂，使用起来不太方便，有时迭代过程只收敛到局部极小解，最坏时甚至不收敛。不过，对于常用的非线性生物数学模型(如指数参数模型，S 型生长模型等)来说，Gauss－Newton 法的迭代过程是收敛的；于是人们退而求其次，大多数生物统计，生物医学，药物动力学方面的文献，均采用变量代换的方法，求出非线性生物模型的回归参数。该方法的优点是应用方便，计算过程简单，有时只需一个带编程的计算器就可解决问题。该方法的不足之处是对原始数据的信噪比(SNR)敏感，此时求出的非线性生物模型的回归参数误差大，严重时甚至会导致求出的非线性生物模型失效。本文重点讨论大多数生物文献中常用的非线性生物数学模型，推导出非线性生物模型回归参数计算的一个新方法。一方面保持了变量代换求非线性生物模型的回归参数的优点；另一方面又有效的解决了变量代换求非线性生物模型的回归参数的缺点，最后给出其应用实例及其分析。

1 新方法的解析说明

在大多数生物文献中应用最广泛的非线性生物数学模型主要有下面几类：

1)指数参数生物数学模型

$$y=\alpha\exp(\beta g(x)) \tag{2}$$

2)幂参数生物数学模型

$$y=\alpha[g(x)]^{\beta} \tag{3}$$

3)生物 S 形生长数学模型

$$y=\frac{1}{\alpha+\beta g(x)} \tag{4}$$

4)特殊幂参数与指数参数生物数学模型

$$y=[g(x)]^{\alpha}\exp[\beta h(x)] \tag{5}$$

其中 α 和 β 是待求参数；$g(x)$ 和 $h(x)$ 是不含参数的任意函数。上面只列出了几个较典型的非线性生物数学模型，实际上下面讨论的方法适合于其它类似的非线性生物数学模型。

如果由实验得到的非线性生物数学模型在 X－Y 空间上的数据集合为：

$\{((x_1,x_2,\cdots x_m)_i,y_i)\mid i=1,2\cdots,n\}$。

在 X－Y 空间上，用如上所列的待拟合的非线性生物数学模型求出的理论数据集合为：

$\{((x_1,x_2,\cdots x_m)_i,\hat{y}_i)\mid i=1,2\cdots,n\}$。

为把 X－Y 空间上的数据转换为 X－Z 空间上的数据，设变量代换

$$z=F(y)。 \tag{6}$$

于是实验数据集合在新变量空间 X－Z 上的映像集合是：

$\{((x_1,x_2,\cdots x_m)_i,z_i)\mid i=1,2\cdots,n\}=\{((x_1,x_2,\cdots x_m)_i,F(y)_i)\mid i=1,2\cdots,n\}$。

理论数据集合在新变量空间 X－Z 上的映像集合是：

$\{((x_1,x_2,\cdots x_m)_i,\hat{z}_i)\mid i=1,2\cdots,n\}=\{((x_1,x_2,\cdots x_m)_i,F(\hat{y}_i))\mid i=1,2\cdots,n\}$。

大多数生物统计，生物医学，药物动力学，化学反应动力学等方面的教科书及论文，在利用以上实验数据求非线性生物数学模型时，总是先求出新变量空间 X－Z 上的残差平方和。

$$S_1=\sum_{i=1}^{n}(z_i-\hat{z}_i)^2。 \tag{7}$$

然后利用最小二乘法求出参数 α 的最佳估计值 $\hat{\alpha}=(\hat{\alpha}_1,\hat{\alpha}_2,\cdots\alpha_l)\in R^l$，代入方程(1)即得所求的非线性生物数学模型。

上面求非线性生物数学模型的过程，隐藏了一个严重缺陷。满足新变量空间 X－Z 上的数据的残差平方和 S_1 最小的参数 α，不一定能满足原变量空间 X－Y 上的数据的残差平方和 S_2 最小。其中

$$S_2=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2。 \tag{8}$$

正是这一缺陷导致了按上面过程求得的非线性生物数学模型回归参数误差大，严重时甚至会导致非线性生物模型失效。

从上面的分析可见，要想导出较理想的非线性生物数学模型，必须使原变量空间 X－Y 上的数据的残差平方和 S_2 最小。

把含未知参数 α 的函数 $\hat{z}_i=F(\hat{y}_i)=F[y(x_i)]$ 在 y_i 处按泰勒(Taylor)级数[1]展开得：

$$(\hat{z}_i)=F(\hat{y}_i)=F(y_i)+F'(y_i)(\hat{y}_i-y_i)+\frac{1}{2}F''(y_i)(\hat{y}_i-y_i)^2+\cdots$$

$$=z_i+F'(y_i)(\hat{y}_i-y_i)+\frac{1}{2}F''(y_i)(\hat{y}_i-y_i)^2+\cdots$$

显然，要导出较理想的非线性生物数学模型，必须使 $\hat{y}_i\to y_i$。此时，去掉上式右边无穷小量$(\hat{y}_i-y_i)$的高阶无穷小后可推得：

$$(\hat{y}_i-y_i)\approx\frac{(\hat{z}_i-z_i)}{F'(y_i)} \tag{9}$$

于是可得原变量空间 X－Y 上的数据的残差平方和 S_2 的近似表达式是：

$$S_2=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2\approx\sum_{i=1}^{n}\frac{(z_i-\hat{z}_i)^2}{[F'(y_i)]^2} \tag{10}$$

其中 $z_i=F(y_i)$，$\hat{z}_i=F[y(x_i)]$，$F'(y_i)=F'(y)|_{y=y_i}$。

对原变量空间 X－Y 上的数据的残差平方和 S_2 的近似表达式应用最小二乘法，求出对应的非线性生物数学模型的正规方程，然后求出参数 α 的最佳估计值 $\hat{\alpha}=(\alpha_1,\alpha_2,\cdots,\alpha_l)\in R^l$，代入方程(1)即得原变量空间 X－Y 上的待求的非线性生物数学模型。

2 应用实例分析

为实际检验新方法的效果，下面给出几个应用

维普资讯 http://www.cqvip.com

实例。为叙述方便,称原来的方法为旧最小二乘近似回归法,本文的方法为新最小二乘近似回归法。

1)如果原始实验数据的信噪比(SNR)较差,旧最小二乘近似回归法实效时,新最小二乘近似回归法仍然能保持较好的回归精度。下面的例 1 摘录自文献[2]。

例 1 根据药学的蛋白结合理论,蛋白的 Langmuir 型单分子层具有吸附功能。在恒温条件下,每克蛋白的药物吸收量 y 与血浆浓度 x 的关系为:

$$y=\frac{\beta x}{\alpha+x} \tag{11}$$

试根据如下 10 组实验数据求给定的非线性生物数学模型中的参数 α 和 β 的估计值。

表 1 原始实验数据

x	12.7	21.2	51.7	77.2	212.4	9.5	22.5	42.3	67.8	243.8
y	0.103	0.466	0.767	1.573	2.462	0.083	0.399	0.899	1.735	2.360

对 10 组实验数据分别用 Gauss - Newton 法,旧最小二乘近似回归法,新最小二乘近似回归法求得的非线性生物数学模型中的参数 α 和 β 的估计值及残差平方和 $\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$ 分别是:

Gauss - Newton 法:

$\hat{\alpha}=144.660$ $\hat{\beta}=4.050$ 残差平方和 $\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=0.4393$

旧最小二乘近似回归法:

$\hat{\alpha}=-103.302$ $\hat{\beta}=-0.865$ 残差平方和 $\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=33.6095$

新最小二乘近似回归法:

$\hat{\alpha}=85.413$ $\hat{\beta}=3.386$ 残差平方和 $\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=0.6988$

从每克蛋白的药物吸收量 y 与血浆浓度 x 的关系式可知,如果血浆浓度 x 趋向于正无穷大时(真实的实验是不可能出现这种情况的),每克蛋白的药物吸收量 y 将趋向于模型中的参数 β。模型中的参数 β 即为每克蛋白的药物吸收量 y 的理论饱和值。根据药学的蛋白结合理论及 10 组实验数据可推出,给定的非线性生物数学模型中的参数 β 至少应当满足的条件是:

$$\beta>\max\{y_i\mid i=1,2,\cdots,n\}=2.462$$

但是用旧最小二乘近似回归法计算出的参数 β 的估计值为 -0.865,与真实的实验过程大相径庭,分析残差平方和数据或观测标准差残图与拟合图可知旧最小二乘近似回归法失效。

用新最小二乘近似回归法计算出的参数 β 的估计值为 3.386,与真实的实验过程基本相符,分析残差平方和数据或观测标准残差图与拟合图可知新最小二乘近似回归法是成功的。尽管新最小二乘近似回归法计算出的参数 β 的估计值与用 Gauss - Newton 法计算出的参数 β 的估计值有些差距,但标准残差图与拟合图及残差平方和数据均能保持基本一致。按照数理统计理论的要求,在血浆浓度 $x\in[9.5,234.8]$ 的实验控制范围内,用新最小二乘近似回归法计算出的参数 α 和 β 的估计值是适用的。最值得一提的是:新最小二乘近似回归法与旧最小二乘近似回归法在计算参数 α 和 β 的估计值时,所用的计算方法与时间基本相同,结果却是前者成功而后者失效。

2)如果不能获得全部原始实验数据,只能得到部分原始实验数据及拟合出的非线性生物数学模型(这在发表的论文中是常见的现象)。分别用旧最小二乘近似回归法与新最小二乘近似回归法对这部分原始实验数据进行拟合,然后对拟合出的两个非线性生物数学模型进行比较,就可以推断出全部原始实验数据的信噪比(SNR),进一步还可以判断出原作者拟合的非线性生物数学模型的优劣。下面的例 2 摘录自文献[3]。

例 2 表 2 中的自变量 Kow 是辛醇/水分配系数,自变量 x 是一阶项分子链接性指数,因变量 EC_{50} 是使发光菌减少一半时的水相有机物浓度值,可以用毒性测量仪测定。文献[3]只给出因变量在新变量空间上的转换数据 $\log(1/EC_{50})$,表 2 中的因变量 EC_{50} 一栏的数据是根据文献[3]中给出新变量空间上的转换数据 $\log(1/EC_{50})$ 换算出来的。

文献[3]选择的非线性生物数学模型是:

$$EC_{50}=\alpha\cdot(K_{OW})^{\beta}\cdot\exp(\kappa x) \tag{12}$$

文献[3]用旧最小二乘近似回归法拟合的非线性生物数学模型是:$EC_{50I}=\exp[-0.912x-0.113\log(K_{OW})-0.480]$ (13)

维普资讯 http://www.cqvip.com

表 2 原始数据的回归分析

LogKow	x	EC_{50}(mol/1000L)	$EC_{50\,I}$	residual	$EC_{50\,II}$	residual	$EC_{50\,III}$	residual
5.05	4.58	4.07	5.37	−1.32	5.31	−1.27	6.66	−2.61
4.27	4.07	10.78	9.33	1.45	8.55	2.23	10.06	0.72
3.59	3.52	12.40	16.64	−4.24	13.94	−1.54	15.49	−3.09
2.81	3.01	21.07	28.94	−7.87	22.42	−1.35	23.41	−2.34
3.31	3.40	20.65	19.16	1.49	15.81	4.48	17.21	3.44
2.28	3.98	11.80	12.68	−0.89	12.20	−0.41	12.59	−0.80
2.13	2.50	35.44	49.75	−14.32	35.53	−0.10	35.08	0.35
2.55	3.57	15.00	17.88	−2.89	15.67	−0.68	16.31	−1.31
0.98	2.52	37.63	55.63	−18.01	41.48	−3.85	38.03	−0.40
2.90	4.10	11.68	10.60	1.08	10.23	1.44	11.03	0.65
1.46	3.06	26.25	32.20	−5.95	26.38	−0.13	25.27	0.98
$\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$				660.75		51.80		38.26

表 2 中的因变量 $EC_{50\,I}$ 一栏的数据是根据上式计算出来的。根据表 2 中的给出的原始数据，本文用旧最小二乘近似回归法拟合的非线性生物数学模型是：

$$EC_{50\,II}=0.2847(K_{OW})^{-0.1468}\exp(-0.7073x) \quad (14)$$

表 2 中的因变量 $EC_{50\,II}$ 一栏的数据是根据上式计算出来的。根据表 2 中的给出的原始数据，本文用新最小二乘近似回归法拟合的非线性生物数学模型是：

$$EC_{50\,III}=0.2309(K_{OW})^{-0.0820}\exp(-0.6839x) \quad (15)$$

表 2 中的因变量 $EC_{50\,III}$ 一栏的数据是根据上式计算出来的。比较表 2 中因变量 EC_{50} 的三组预测数据及其残差平方和，可以看出非线性生物数学模型(14)的回归精度明显优于非线性生物数学模型(13)，提示文献[3]未给出的实验数据的信噪比(SNR)低，因为非线性生物数学模型(13)和非线性生物数学模型(14)同是由旧最小二乘近似回归法拟合的。非线性生物数学模型(15)的回归精度优于非线性生物数学模型(14)，但不太显著，提示文献[3]给出的这部分实验数据的信噪比(SNR)较高。旧最小二乘近似回归法的弱点未充分暴露，新最小二乘近似回归法的优点未充分发挥作用。这正好可以作为判断实验数据的信噪比(SNR)高低的一条准则。从非线性生物数学模型(15)优于非线性生物数学模型(14)；非线性生物数学模型(14)明显优于非线性生物数学模型(13)可以断定；倘若将本文的新最小二乘近似回归法用于文献[3]中的全部原始实验数据，肯定能得到明显优于文献[3]拟合的非线性生物数学模型(13)的非线性生物数学模型。

3 结束语

通过上面的讨论可知，新旧最小二乘近似回归法相同之处是：均引进了变量代换 $z=F(y)$；不同之处是：旧最小二乘近似回归法利用新变量空间 X－Z 上的残差平方和 $S_1=\sum_{i=1}^{n}(z_i-\hat{z}_i)^2$，然后使用最小二乘法求出参数 α 的最佳估计值 $\hat{\alpha}=(\alpha_1,\alpha_2,\cdots\alpha_l)\in R^l$，代入方程(1)得到所求的非线性生物数学模型；而新最小二乘近似回归法则是利用原变量空间 X－Y 上的数据的残差平方和 $S_2=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$，然后使用最小二乘法求出参数 α 的最佳估计值 $\hat{\alpha}=(\alpha_1,\alpha_2,\cdots\alpha_l)\in R^l$，代入方程(1)得到所求的非线性生物数学模型。正是这点差别使新最小二乘近似回归法所求的非线性生物数学模型比旧最小二乘近似回归法所求的非线性生物数学模型显著地提高了回归精度。另外，新最小二乘近似回归法所导出的正规方程仅是旧最小二乘近似回归法所导出的正规方程的加权式，保留了基使用方便的优点。

参 考 文 献

[1] 张世强，主编. 医学高等数学[M]. 北京：科学出版社，2001；250－251.

[2] 方开泰，全 辉，陈庆云. 实用回归分析[M]. 北京：科学出版社，1988；78.

[3] 赵元慧，高 鸿. 有机污染物定量结构与活性相关性分析[J]. 科学通报，1993；38(1)：30－32.