局部模型回归分析

胡良平[1,2*]

(1. 军事科学院研究生院,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】 本文目的是介绍局部模型回归分析的概念、作用以及如何用软件实现计算的方法。先介绍有关的基本概念,再介绍基本原理,最后通过一个实例并基于 SAS 软件演示如何实施局部模型回归分析。结果表明:局部模型回归分析最适合用于“全部观察点呈现线性递增或下降趋势”,在多个“小区域或邻域”上表现为“二次曲线”或“三次曲线”形状或具有某种“聚集性”的场合。

【关键词】 局部模型回归分析;光滑参数;二次曲线;三次曲线
中图分类号:R195.1 文献标识码:A doi:10.11886/j.issn.1007-3256.2018.04.003

Local model regression analysis

Hu Liangping[1,2*]

(1. *Graduate School, Academy of Military Sciences PLA China, 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 was to introduce the concepts and functions as well as the calculation methods by using the statistical software of the local model regression analysis. Firstly, the basic concepts of the regression analysis were introduced. Secondly, the basic principles of the regression analysis were given. Finally, the local model regression analysis was demonstrated through one example by using the SAS software. The results showed that it was suitable for using the method mentioned above when there were some features in the data, such as the linear trend in ascending or descending form with the quadratic curves or cubic curves or being a certain "clusters" in multiple intervals or neighborhoods of the independent variable.

【**Keywords**】 Local model regression analysis; Smooth parameter; Quadratic curve; Cubic curve

1 概 述[1]

1.1 局部回归模型

局部回归模型见式(1):

$$y_i = g(x_i) + \varepsilon_i \qquad i = 1,2,\cdots,n \qquad (1)$$

在式(1)中,y_i 为第 i 次观测到的因变量的取值;$g(x_i)$ 是 x_i 的回归函数;x_i 可以是一个自变量,也可以是由多个自变量组成的向量;ε_i 是一个随机误差。

1.2 局部模型回归分析应用的场合

一般来说,在因变量服从正态分布或对称分布时,欲研究因变量随自变量变化而变化的依赖关系时,可以尝试采用很多种方法来创建回归模型,包括采用“局部回归模型”。最适合运用此模型的场合如下:在自变量的全部取值范围内,存在多个“小区域”,在这些“小区域”内,观测点的密度较高,似乎呈现出“聚集性”;而且,它们或呈“二次多项式曲线形状”或呈“三次多项式曲线形状”分布。见图 1。

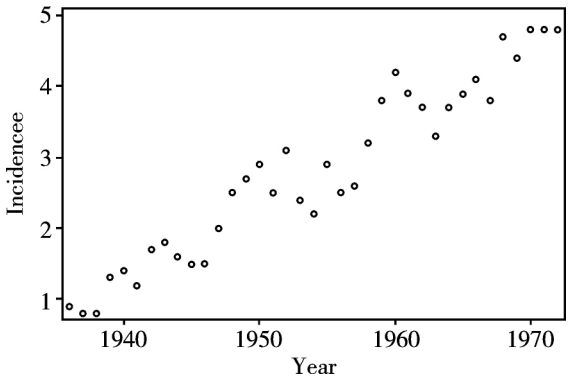

图 1 黑色素瘤发病率随时间推移的变化趋势

1.3 局部模型回归分析的计算原理

1.3.1 计算原理

所谓局部模型,实际上就是在每个“小区域或小邻域”上构建自变量的一个线性或二次曲线模型、甚至三次曲线模型。问题在于如何选取一系列的“小邻域”。一个最直观的想法是:将全部数据观察点按自变量由小到大的顺序排列,先确定由多少

项目基金:国家高技术研究发展计划课题资助(2015AA020102)

个相邻的观察点决定一个“小邻域”，比如，设观察点数目为 $k(k\geqslant 3)$，当 k 取一个确定数值后，就很容易将全部观察点划分成 m 个“小邻域”。于是，在每个“小邻域”上创建一个“局部模型”，计算出各“小邻域”上因变量的残差平方和，再求出所有“小邻域”上残差平方法和之和，就可获得总残差平方和。接下去，就可以改变 k 值，假定令 $k=3$ 到 $k=n$（即全部观察点）共有 j 种情况，由前面的计算就可获得某种情况下的“总残差平方和”最小，于是，就认为按这种情况对应的“k 值”来形成“小邻域”是最合适的。

事实上，在 SAS 的 LOESS 过程中，评价拟合效果所选用的统计量为校正的赤池信息准则（AICC）（其取值越小越好，具体计算公式详见后文），它所对应的 k 值被转换成“光滑参数 s”，s = k/n（其中 k 需要事先依据某种方法或理由初步估计出来，n 为样本含量或全部观察点数目）。在每个“小邻域”上建模时，采用“加权最小平方法”[2]。

1.3.2 常用的拟合效果评价指标

（1）赤池信息准则（The Akaike information criterion，AIC）：AIC 是模型对资料拟合优度的一种度量，也体现了现在所使用的模型相对于最简约模型之间的一种平衡。其定义如下：

$$AIC = -2LL + 2p$$

上式中，p 为模型中被估计参数的个数，LL 是用于估计参数数值的似然函数的对数。

（2）AICC：

$$AICC = -2LL + 2p\frac{n}{n-p-1}$$

上式中，n 为总样本含量，其他变量含义同上。

（3）贝叶斯信息准则（Bayesian Information Criterion，BIC）与 AIC 和 AICC 是类似的度量，其定义如下：

$$BIC = -2LL + p\log(n)$$

上式中，各变量的含义同上，此处不再赘述。

2 基于局部模型回归分析解决实际问题[1]

2.1 问题与数据结构

【例 1】下面是一个关于黑色素瘤发病率的资料。资料来自美国康涅狄格州肿瘤注册部门，时间从 1936 年 - 1972 年共 37 年，基于年龄校正的各年黑色素瘤的发病率（1/10 万）的前 8 年数据见表 1，其他数据详见后面的 SAS 程序：

表 1 基于年龄校正的 1936 年 - 1943 年黑色素瘤发病率

年 份	发病率	年 份	发病率
1936 年	0.9	1940 年	1.4
1937 年	0.8	1941 年	1.2
1938 年	0.8	1942 年	1.7
1939 年	1.3	1943 年	1.8

【对数据结构的分析】严格地说，这是一个“时间序列”数据，即发病率随着时间的推移而动态变化。为简便起见，暂且将该数据视为一个计量因变量 y（发病率）随另一个计量自变量 x（年份）变化的依赖关系问题。

【统计分析方法的选择】研究 y 与 x 之间依赖关系的最简单方法是进行直线回归分析；若两变量之间呈曲线变化趋势，就可选择某种曲线方程进行曲线回归分析。

2.2 基于常规方法构建简单线性回归模型[3]

2.2.1 创建 SAS 数据集

创建一个名为“melanoma”的临时 SAS 数据集的 SAS 数据步程序如下：

```
data Melanoma;
input Year Incidences @@;
format Year d4.0;
datalines;
 1936  0.9  1937  0.8  1938  0.8  1939  1.3
 1940  1.4  1941  1.2  1942  1.7  1943  1.8
 1944  1.6  1945  1.5  1946  1.5  1947  2.0
 1948  2.5  1949  2.7  1950  2.9  1951  2.5
 1952  3.1  1953  2.4  1954  2.2  1955  2.9
 1956  2.5  1957  2.6  1958  3.2  1959  3.8
 1960  4.2  1961  3.9  1962  3.7  1963  3.3
 1964  3.7  1965  3.9  1966  4.1  1967  3.8
 1968  4.7  1969  4.4  1970  4.8  1971  4.8
 1972  4.8
;
run;
```

2.2.2 绘制散布图，直观展示两变量之间的变化趋势

利用下面的 SAS 过程步程序，可以绘制反映两变量变化趋势：

```
proc sgplot data = Melanoma;
scatter y = Incidences x = Year;
run;
```

【SAS 输出结果】

第 1 部分输出结果为“图 1”，已经在前面呈现，此处从略。

由图 1 可看出：散点呈上升的变化趋势。但仔细观察散点，发现在多个局部区域内散点表现为“聚集性”，并且呈“矩形”或“三角形”等形状。

下面尝试采用简单直线回归模型拟合该资料：

```
ods graphics on;
proc reg data = Melanoma;
model Incidences = Year;
run;
```

【SAS 主要输出结果】

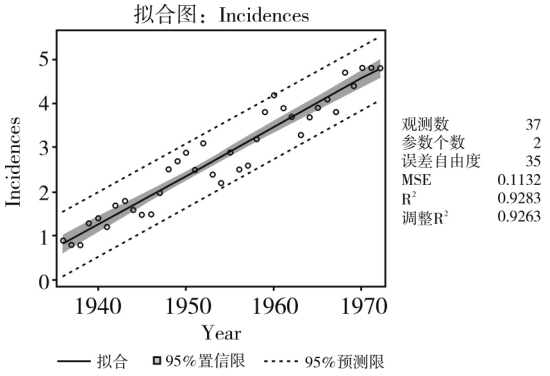

图 2 采用直线回归模型描述黑色素瘤发病率随时间推移的变化趋势

拟合的统计量：均方根误差 = 0.33641、R^2 = 0.9283、调整 R^2 = 0.9263，从这些拟合统计量的数值来看，似乎用简单直线回归模型拟合此资料效果相当令人满意。但从图 2 可看出：在多个局部区域上，直线不能很好地给出预测结果。

2.3 基于局部模型构建非线性回归模型[1]

基于局部模型构建非线性回归模型的 SAS 程序如下：

```
proc loess data = Melanoma;
model Incidences = Year;
run;
```

【SAS 程序说明】以上 SAS 程序调用 LOESS 过程拟合局部模型。

【SAS 输出结果及其解释】

由图 3 可看出：局部模型对此资料的拟合效果非常好，既没有“过拟合”，也没有“欠拟合”。

如何才能做到既不“过拟合”又不“欠拟合”？关键是要选取合适的“光滑参数”，它已显示在图 3 的左上角，即“Smooth = 0.257”。用此数值乘以总样本含量 37 等于 9.5，说明程序按横坐标轴的顺序，将每相邻 9 或 10 个观测点所在的区域视为一个“局部区域”，在该区域上进行多项式拟合。

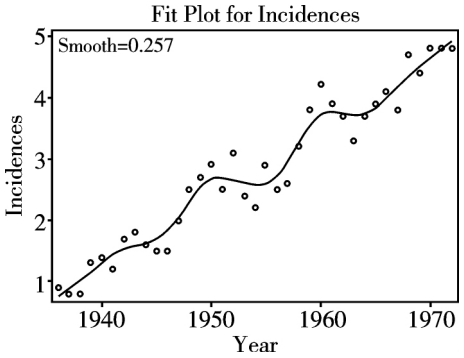

图 3 采用局部模型拟合的结果

如何获得最佳“光滑参数”的数值？在 SAS 的 LOESS 过程中，先给定一系列的“光滑参数”值进行拟合，对于每个给定的“光滑参数”值，就能计算出若干个反映拟合效果或优度的统计量，其中，以 AICC 统计量取得最小值时对应的“光滑参数”为最佳。

利用如下 SAS 程序可以同时获得 4 个“光滑参数”对应的拟合结果，

```
proc loess data = Melanoma plots = ResidualsBySmooth
( smooth) ;
model Incidences = Year/smooth = 0.1 0.25 0.4 0.6;
run;
```

【SAS 主要输出结果】

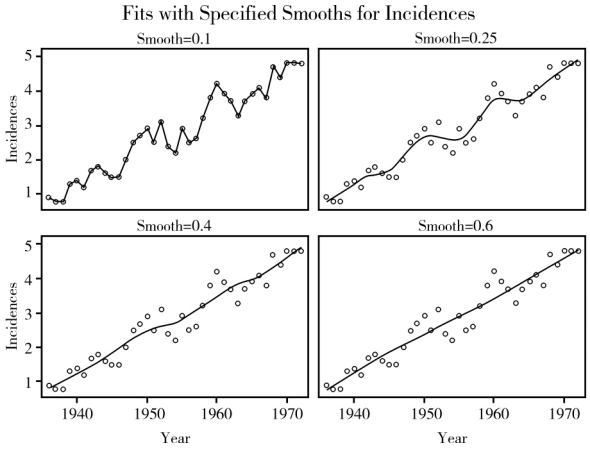

图 4 基于 4 个光滑参数进行局部模型拟合得到的拟合结果

在图 4 中有 4 幅小图，从上往下、从左往右的“光滑参数”依次为 0.1、0.25、0.4 和 0.6 对应的拟合结果。不难看出：“Smooth = 0.1”属于“过拟合”，而“Smooth = 0.4”和“Smooth = 0.6”属于“欠拟合”，只有“Smooth = 0.25”，属于“正常拟合”，因为它已经是最佳“光滑参数”0.257 的近似值。

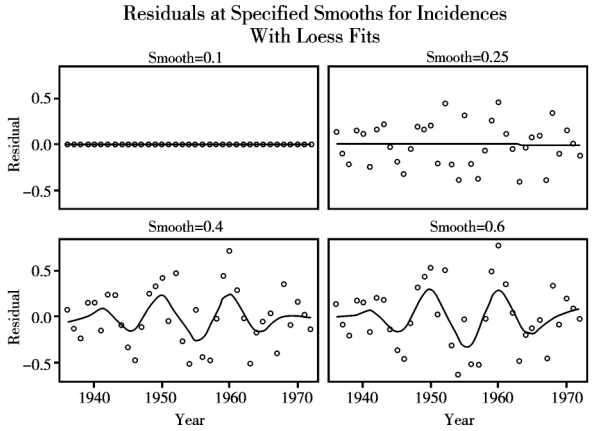

图 5 基于 4 个光滑参数进行局部模型拟合得到的残差图

图 5 中的 4 幅小图分别与图 4 中 4 幅小图一一对应，只不过图 5 反映的是残差。当“Smooth = 0.1”时，几乎所有观察点上的残差都为 0，这就是“过拟合”；当“Smooth = 0.25”时，残差图上散点在各处波动接近且没有明显的变化趋势，属于“正常拟合”；而图 5 中下面的 2 幅小图都呈现出残差散点具有一定的变化规律，属于“欠拟合”。

为了避免盲目性，可以采用下面的 SAS 程序自动寻找到最佳的“光滑参数”的数值：

```
proc loess data = Melanoma;
model Incidences = Year / details( ModelSummary OutputStatistics) ;
run;
```

【SAS 主要输出结果】

Model Summary

Smoothing Parameter	Local Points	Residual SS	GCV	AICC
0.41892	15	3.42229	0.00339	−0.96252
0.68919	25	4.05838	0.00359	−0.93459
0.31081	11	2.51054	0.00279	−1.12034
0.20270	7	1.58513	0.00239	−1.12221
0.17568	6	1.56896	0.00241	−1.09706
0.28378	10	2.50487	0.00282	−1.10402
0.20270	7	1.58513	0.00239	−1.12221
0.25676	9	2.03105	0.00252	−1.17277
0.22973	8	2.02965	0.00256	−1.15145
0.25676	9	2.03105	0.00252	−1.17277

以上是程序自动寻找最佳“光滑参数”的动态过程，仅当局部观测点为 9 个时，AICC 统计量能取到最小值 −1.17277，此时，对应的“光滑参数”为 0.25676。

Fit Summary

Fit Method	kd Tree
Blending	Linear
Number of Observations	37
Number of Fitting Points	37
kd Tree Bucket Size	1
Degree of Local Polynomials	1
Smoothing Parameter	0.25676
Points in Local Neighborhood	9
Residual Sum of Squares	2.03105
Trace [L]	8.62243
GCV	0.00252
AICC	−1.17277

以上是模型拟合效果的总结。

利用下面的 SAS 程序，可以得到拟合曲线的置信带：

```
proc loess data = Melanoma;
model Incidences = Year/clm alpha = 0.05;
run;
```

【SAS 主要输出结果】

Fit Summary

Fit Method	kd Tree
Blending	Linear
Number of Observations	37
Number of Fitting Points	37
kd Tree Bucket Size	1
Degree of Local Polynomials	1
Smoothing Parameter	0.25676
Points in Local Neighborhood	9
Residual Sum of Squares	2.03105
Trace [L]	8.62243
GCV	0.00252
AICC	−1.17277
AICC1	−42.03789
Delta1	27.06596
Delta2	26.76564
Equivalent Number of Parameters	7.31083
Lookup Degrees of Freedom	27.36964
Residual Standard Error	0.27394

以上是模型拟合效果的总结，与前面给出的结果基本相同。

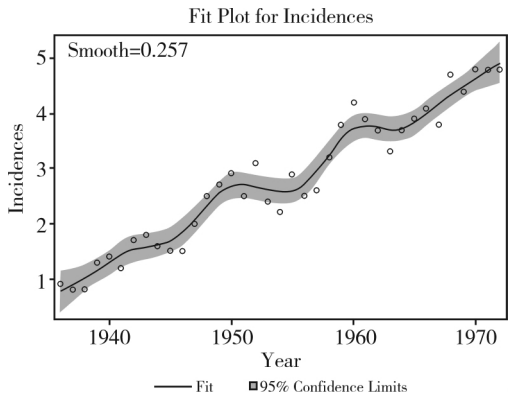

图 6 基于光滑参数为 0.257 时得到的局部多项式拟合结果及 95% 置信带

2.4 小结

从上面的介绍可知: 局部模型的关键在于选取“光滑参数”的具体取值。此值的真实含义是以每相邻的多少个观察点为一个“小区域”,在每个这样的“小区域”上拟合一个“多项式”。当“Smooth = 0.1”(相当于样本含量的 1/10 的观察点)时,得到了“过拟合”的结果。就本例而言,37/10 = 3.7 ≈4,若采用 4 次多项式,则多项式曲线就会通过每个观察点; 当“Smooth = 0.6”(相当于样本含量的 6/10 的观察点)时,得到了“欠拟合”的结果。就本例而言,6 × (37/10) ≈22,若采用 4 次多项式,则多项式曲线就很难通过大多数观察点。

当采用简单直线回归模型时,就相当于取“Smooth = 1.0”,也就把全部观察点所在的范围视为一个“小区域”,采用一个“一次多项式”去拟合资料,这对于具有类似图 1 中散点所表现的状态是没有任何帮助的。

由此可知: 局部模型最适合用于如下的资料: 全部观察点呈现线性递增或下降趋势,而在多个“小区域”上表现为“二次曲线”或“三次曲线”或“四次曲线”的形状。建模的目的只是为了形象化地拟合数据并对未知因变量的取值进行预测,而不需要呈现回归模型的具体表达式(因此法不便给出具体的回归模型)。

参考文献

[1] SAS Institute Inc. STAT SAS 9.3 User's Guide [M]. Cary, NC: SAS Institute Inc, 2011: 3965 - 4032.
[2] 胡良平,胡纯严,鲍晓蕾. 应用数理统计 [M]. 北京: 电子工业出版社, 2015: 142 - 192.
[3] 谷恒明, 胡良平. 简单线性回归分析及其应用 [J]. 四川精神卫生, 2017,30(6): 494 - 497.

(收稿日期: 2018 - 08 - 10)
(本文编辑: 陈 霞)