·科研方法专题·

变量变换回归分析(Ⅰ)
——拟合含间断点资料的方法

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】 本文目的是介绍一种能很好地拟合具有间断点资料的方法。当资料中具有明确的间断点或整个资料包含多段不同变化趋势的曲线类型时,为了提高曲线回归模型对资料的拟合优度,需要充分发挥“节点”的作用。可基于两种不同视角来利用“节点”:其一,人为设定不同数目的节点,利用样条变换方法拟合分段多项式曲线;其二,在客观存在的节点上,求曲线的一阶乃至四阶导数,并据此构建曲线回归模型。得到的结论是:后者的拟合效果优于前者。

【关键词】 节点;样条;变量变换;曲线拟合;回归分析
中图分类号:R195.1 文献标识码:A doi:10.11886/j.issn.1007-3256.2019.03.001

Regression analysis based on the variable transformation(Ⅰ)
——the approach of fitting the data with discontinuous points

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 paper introduced one good approach of fitting the data with discontinuous points. In order to improve the goodness of curve fitting, it was necessary that the knots should play a role when the following situations appeared, such as the data with definite discontinuous points and/or the data with several types of curvilinear trend in the whole interval of the independent variable. From two perspectives, the knots could be deal with the two approaches as follows. First, artificial setting various number of the knots and then fitting the piece-wise polynomial curve by using the spline transformation. Second, the first derivative, the second derivative, until the fourth derivative was computed on the objectively existent knots, based on the previous results, the curve regression model was structured. The conclusion proved that the fitted effect of the latter approach provided better fitting effect than the former.

【**Keywords**】 Knot; Spline; Variable transformation; Curve fitting; Regression analysis

1 基本概念

1.1 变量变换

在回归分析中,常需要对变量进行变换,以使资料满足待拟合回归模型的要求。最常见的情形是对定性自变量作哑变量变换,对定量变量作对数变换、指数变换、平方根变换或平方变换等。根据变量变换所涉及的原理和计算方法的不同,在 SAS 的 TRANSREG 过程中,变量变换方法分为五大类,每一类中又可以细分成许多种小类[1]。因篇幅所限,此处暂不赘述。

1.2 样条变换

“样条变换”是变量变换中颇具特色的一类变换方法,其基本思路是拟合“分段多项式曲线”。若根据具体算法,样条变换可以进一步划分为“B-样条变换”“B-样条基函数变换”“单调 B-样条变换”“无迭代惩罚 B-样条变换”“迭代光滑样条变换”和“非迭代光滑样条变换”等子类[1]。

1.3 节点

所谓“节点”,就是“间断点”或“不连续点”。通常有两种“节点”:其一,实际资料中客观存在的“间断点”,例如本文的图 1 中就出现了 3 个“节点”;其二,统计分析者人为设定的“间断点”,例如根据某实际资料的散布图中各散点的分布情况,似乎可以分别采用“对数曲线”“指数曲线”和“幂函数曲线”去描述该实际资料的变化趋势,于是,分析者就需要在自变量的取值区间内设定两个“节点”,将

项目基金:国家高技术研究发展计划课题资助(2015AA020102)

整个曲线划分成三段。对此资料的拟合优度高低，取决于两个“节点”的位置选得是否恰当。本文将针对前述提及的两种“节点”分别拟合曲线，以说明在进行曲线拟合时，如何发挥“节点”的作用。

2 一个人工生成的实例——由三个间断点分割的四段曲线

2.1 产生实例数据集的 SAS 程序

利用下面的 SAS 程序，可以生成一个由三个间断点分割的四段曲线。

```
title1'An Illustration of Splines and Knots';
* Create in y a discontinuous function of x.;
data a;
x = -0.000001;
do i=0 to 199;
if mod(i, 50) = 0 then do;
c = ((x/2) -5) * *2;
if i=150 then c = c +5;
y = c;
end;
x = x +0.1;
y = y - sin(x - c);
output;
end;
run;
```

2.2 数据集 a 的基本情况描述与呈现

数据集 a 中主要有两个计量变量，即 x 与 y，共有 200 个观测点。x 与 y 的简单统计量如下。

变量	N	均值	标准差	最小值	最大值
x	200	10.0499990	5.7879185	0.0999990	19.9999990
y	200	12.0433539	9.2339544	-7.3705670	29.1657155

其中，x 呈均匀分布、y 呈负偏态分布（图形从略）。（x，y）的散布图见图 1。

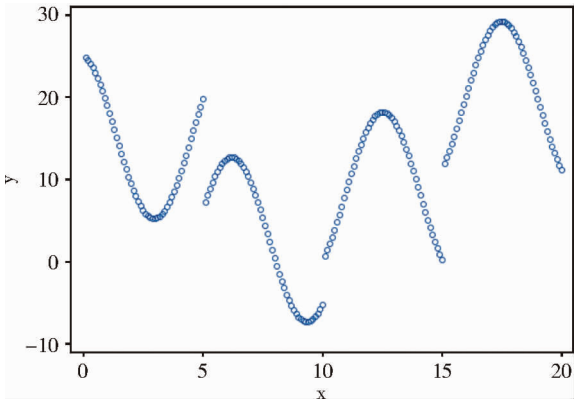

图 1 基于数据集 a 绘制出（x，y）的散布图

由图 1 可知，在变量 x 的整个变化范围[0.0999990，19.9999990]中，共有 4 段曲线，它们被 3 个不连续点（x = 5、10、15）分隔开。从左至右，第一、三和四段曲线很像二次抛物线，第二段曲线很像三次抛物线。

2.3 统计分析的任务

试以 x 为自变量、y 为因变量，建立 y 依赖 x 变化而变化的回归模型。换言之，要对图 1 所代表的数据集进行曲线拟合，希望构建一个回归方程或回归模型，以便用区间[0.0999990，19.9999990]中的任何一个 x 的取值代入，求出因变量 y 的估计值。要求是预测误差应尽可能小。

3 不考虑与考虑节点且使用多项式函数进行曲线拟合

3.1 不考虑节点且仅使用多项式函数进行曲线拟合

3.1.1 呈现各种情况下的拟合效果

所谓“不考虑节点”，就是将全部数据集视为一个“整体”，拟合一条光滑的曲线回归方程或模型。通常，可以考虑从拟合 2 次多项式开始，逐渐增加“次数”。以下依次拟合 2 ~ 16 次多项式，以“R^2”反映其拟合优度。拟合效果见表 1。

表 1 对数据集 a 拟合 2 ~ 16 次多项式所得的拟合效果

编号	多项式的次数	R^2	编号	多项式的次数	R^2
1	2	0.40720	9	10	0.88232
2	3	0.46884	10	11	0.91857
3	4	0.52785	11	12	0.94481
4	5	0.58086	12	13	0.94524
5	6	0.58815	13	14	0.96258
6	7	0.65903	14	15	0.96266
7	8	0.67254	15	16	0.96475
8	9	0.86077			

由表 1 可知，随着多项式“次数”的增加，R^2 也不断增大。但是，多项式的次数每增加 1 次，就相当

于增加了一个新变量。一个包含 16 次多项式的回归模型中包含了 17 个未知参数(因为还包含 1 个常数项),显然增加了模型的复杂程度,模型的实用性大大降低了。不仅如此,模型在“不连续点(x = 5、10、15)”附近的拟合误差很大。见图 2。

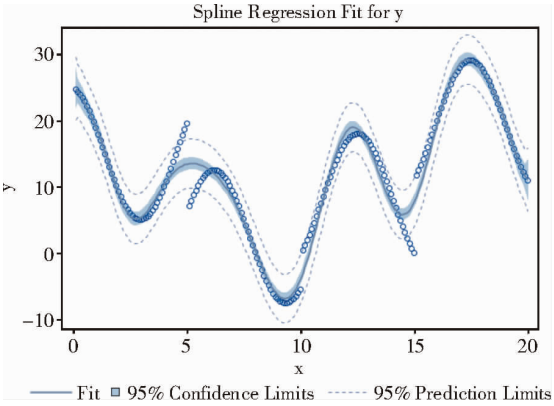

图 2 基于样条变换且采用 16 次多项式拟合数据集 a 的效果

3.1.2 多项式次数为 16 时对应的 SAS 过程步程序

表 1 中最后一行(多项式的次数为 16)所对应的 SAS 程序如下:

```
proc transreg data = a;
model identity(y) = spline(x / degree = 16);
run;quit;
```

依次将上述程序中的选项“degree = ”赋值为 2 ~ 15,可获得表 1 中其他各行的计算结果。

3.2 由程序确定节点位置且使用样条变换进行曲线拟合

3.2.1 呈现各种情况下的拟合效果

分析者可以指定节点的个数,一旦个数确定,将由程序按照规定的方式计算出各节点所在的位置。于是,整个数据集就被分割为若干个子区间,程序将在每个子区间上拟合样条函数,这就是所谓的“分段拟合”,分段的数目为节点数加 1。在每一段上,又可以拟合 2、3、4 等高次多项式。但通常的样条函数为 3 次多项式,即分段拟合 3 次多项式。下面仅改变节点数,但在每段上都将拟合 3 次多项式。节点数从 1 到 15,对应的拟合效果见表 2。

由表 2 可知,从总趋势上来看,随着节点数增加,R^2 也在增大,但当节点数为 5、8 和 11 时,R^2 略有下降。由于节点数越多,整个区间被划分成的子区间数目就越多,也即 3 次多项式的个数越多,回归模型越复杂。下面给出当 R^2 取得最大值0.97508(节点数为 14)时,对应的曲线拟合结果。见图 3。

表 2 考虑不同节点数且仅拟合 3 次多项式对数据集 a 拟合的效果

编 号	节点数	R^2	编 号	节点数	R^2
1	1	0.53585	9	9	0.95256
2	2	0.55391	10	10	0.96674
3	3	0.61465	11	11	0.95835
4	4	0.74450	12	12	0.96231
5	5	0.68823	13	13	0.97174
6	6	0.88890	14	14	0.97508
7	7	0.93608	15	15	0.96872
8	8	0.92566			

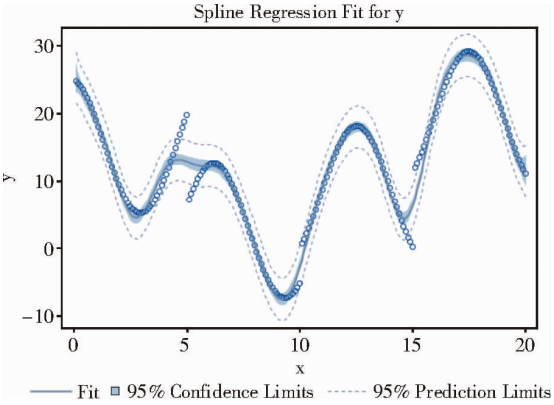

图3 基于节点数为 14 且采用 3 次样条拟合数据集 a 的效果

3.2.2 节点数为 15 时对应的 SAS 过程步程序

表 2 中最后一行所对应的 SAS 程序如下:

```
proc transreg data = a;
model identity(y) = spline(x/degree = 3 nknots = 15);
run;
```

依次将上述程序中的选项“nknots = ”赋值为 1 ~ 14,可获得表 2 中其他各行的计算结果。

3.3 基于数据集的真实情况给定 3 个节点位置且使用 3 次样条变换进行曲线拟合

3.3.1 每个节点出现 k(k = 1 ~ 4)次

在 x = 5、10、15 三处真实的“不连续点”处,对函数求一阶导数;若将三个不连续点各重复写 2 次,就是对函数求二阶导数;同理,若将三个不连续点各重复写 k(k = 1 ~ 10)次,就是对函数求 k 阶导数。求不同阶导数后,拟合的效果可能会有所变化。经尝试,当 k≥4 时,拟合的效果几乎相同。见表 3。

表 3 将三个“不连续点”各重复 k 次且仅用 3 次多项式对数据集 a 拟合的效果

编 号	重复次数(k)	R^2
1	1	0.61730
2	2	0.87941
3	3	0.95542
4	≥4	0.99254

由表 3 可知,将三个“不连续点”各重复 4 次,就能获得最好的拟合效果。见图 4。

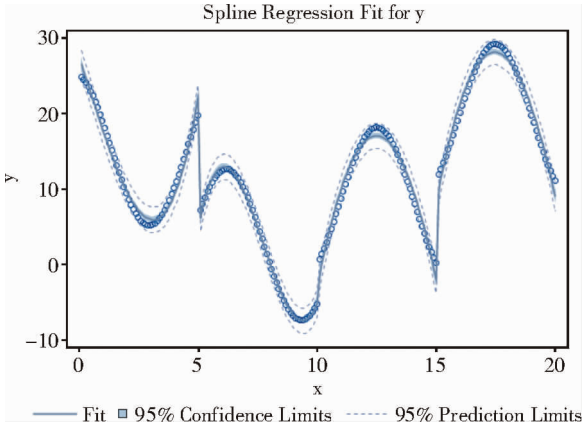

图 4 将三个“不连续点”各重复 4 次且仅用 3 次多项式的拟合效果

3.3.2 将三个“不连续点”各重复 4 次时对应的 SAS 过程步程序

```
proc transreg data = a;
model identity(y) = spline(x / knots =
                           5  5  5  5
                           10 10 10 10
                           15 15 15 15);
run;quit;
```

4 讨论与小结

常规的曲线拟合方法不适合含有间断点的资料,其主要原因在于:常规的曲线类型基本上都在自变量的定义域范围内是连续的[2-4]。本文数据集 a 含有 3 个间断点,故常规的曲线拟合方法不适合。

本文通过三种方法来拟合含有间断点的资料。第一种:不考虑节点且仅考虑多项式的次数,由表 1 可知,欲使 R^2 达到 0.96 以上,至少需要采用 14 次多项式;第二种:基于 3 次多项式且考虑节点的数目,由表 2 可知,欲使 R^2 达到 0.96 以上,至少需要采用 10 个节点;第三种:基于 3 次多项式、瞄准资料中真实的 3 个“间断点”,并强调在“间断点”处重复不同次数,欲使 R^2 达到 0.96 以上,只需要在 3 个“间断点”处重复 4 次即可。

欲达到相同或相近的拟合效果,第一种方法所对应的回归模型过于复杂且实用性很差;第二种方法所对应的回归模型有了很大的改进,回归模型的实用性有了较大提升;第三种方法所对应的回归模型最简洁,也最具有实用性。

参考文献

[1] SAS Institute Inc. STAT SAS 9.3 User's Guide[M]. Cary, NC: SAS Institute Inc, 2011: 7761 - 8002.
[2] 胡良平, 高辉. 非线性回归分析[M]. 北京:电子工业出版社, 2013: 24 - 37, 71 - 116.
[3] 谷恒明, 胡良平. 简单曲线回归分析及其应用[J]. 四川精神卫生, 2017, 30(6): 498 - 502.
[4] 谷恒明, 胡良平. 复杂曲线回归分析及其应用[J]. 四川精神卫生, 2017, 30(6): 503 - 506.

(收稿日期:2019 - 06 - 12)
(本文编辑:陈 霞)

科研方法专题策划人——胡良平教授简介

胡良平,男,1955 年 8 月出生,教授,博士生导师,曾任军事医学科学院研究生部医学统计学教研室主任和生物医学统计学咨询中心主任、国际一般系统论研究会中国分会概率统计系统专业理事会常务理事和北京大学口腔医学院客座教授;现任世界中医药学会联合会临床科研统计学专业委员会会长、中国生物医学统计学会副会长,《中华医学杂志》等 10 余种杂志编委和国家食品药品监督管理局评审专家。主编统计学专著 48 部,参编统计学专著 10 部;发表第一作者学术论文260余篇,发表合作论文 130 余篇,获军队科技成果和省部级科技成果多项;参加并完成三项国家标准的撰写工作;参加三项国家科技重大专项课题研究工作。在从事统计学工作的 30 年中,为几千名研究生、医学科研人员、临床医生和杂志编辑讲授生物医学统计学,在全国各地作统计学学术报告 100 余场,举办数十期全国统计学培训班,培养多名统计学专业硕士和博士研究生。近几年来,参加国家级新药和医疗器械项目评审数十项、参加 100 多项全军重大重点课题的统计学检查工作。归纳并提炼出有利于透过现象看本质的“八性”和“八思维”的统计学思想,独创了逆向统计学教学法和三型理论。擅长于科研课题的研究设计、复杂科研资料的统计分析与 SAS 实现、各种层次的统计学教学培训和咨询工作。