DOI:10.13546/j.cnki.tjyjc.1987.03.022



# 再介绍一种新的回归方法

## ——分段线性回归

· 倪加勋

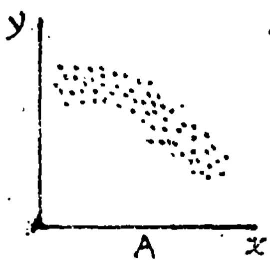
在1986年第3期《统计与决策》杂志上，我曾经介绍过一种单调回归的方法。近年来随着计算机科学的发展，国外数理统计中应用回归这一分支也得到很快的发展。例如，多元回归，过去由于计算工作量比较大，在应用上受到了一定的限制，而现在则已有大量的统计软件可供使用，而且有许多软件如minitab，spss等已经可用于微机，这些软件使用方便，并不需要掌握某种程序语言，只要学会几个简单的指令，把数据输入，很快能得到计算结果，因此回归方法的应用也日益广泛。此外除了传统的回归方法以外，还提出了一些新的方法，可以把一些比较复杂的问题设法转化成比较简单的方法。这里要介绍的一种方法也是近年来新提出的，叫作分段线性回归法（piecewise Linear regression）。

### 一、什么是分段线性回归和它的提出

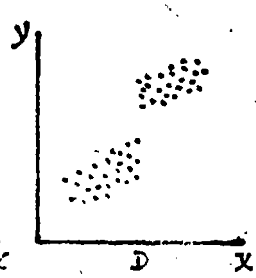
因为客观现象之间的联系总是很复杂的，常常不可能是简单直线关系，因而使用线性回归的方法就受到限制，若使用非线性方法，即使手头有计算功能很大的计算机可供使用，但是用什么曲线去拟合总是要人来判断决定，而这也不是很容易的，目前传统的方法是根据经验来判断，然后在计算机上试行拟合，并计算其残差，要使残差达到最小或使确定系数$r^2$的值尽可能大(趋近于1)为好。然而这种方法也并不常常有效。原因是：（1）曲线的形式比较难找；(2)即使能找出一些特殊的曲线函数形式，往往没有现成的软件。但我们知道使用回归方法的本质就是要通过所取得的资料中变量之间的关系，从中汲取有用的信息为我们所用。这些信息有时可以通过直观的观察取得。我们常用散布(点)图就是直观提取信息的一种重要方法。现在我们通过二个图例来说明分段线性回归方法的提出，设因变量Y，与自变量X之间散布(点)，参图(1)和图(2)。

显然，图(1)和图(2)均不呈直线关系，图(1)为一下凹的曲线，但曲线的形式尚不能肯定，图(2)则是一个不连续的函数，使用传统的方法就比较困难，但是以图上看来我们可以发现二个图形有一个共同的地方，在图(1)自变量X以A为分界点，把数据分成两段，这二段分别呈线性关系；在图(2)中以D点为分界点也把数据分成二段，前后也分别呈线性关系。因此这就启发我们若能区别不同阶段分别用线性回归方法来进行拟合和进行回归分析，也许能起到较好的效果，而且方法也比较简单，因为解线性回归一般均有现成的计算机软件，即使是多元回归也比较容易计算。所以分段线性回归方法实际上就是设法把一比较复杂的曲线，把它分段割成线性的处理方法。

图(1)




图(2)




### 二、分段线性回归的具体进行方法

我们了解了分段回归的基本思路以后，当遇到这种情况时，当然可以把搜集到的数据分成几段，然后分别使用简单线性回归的方法，但是这样做有两个缺点，一是要把数据重新整理，有时也比较麻烦；二是把一个总的样本分割成几个样本后往往样本的容量太小，因而每个子样本的自由度很小，在对回归系数假设检验时，往往是不显著的，分段线性回归方法是借助于多元回归分析中的指示变量，（有的书上也称虚拟变量）使之成为一个多元线性回归，这样就有统计软件可供使用。如果只有分成二段，则一般情况下只有二个自变量和一个因变量的线性回归，借助于普遍的计算器也不难计算，现在分几种情况来谈：

1、有一个转折点的分段线性回归。以图(1)的情况为例，回归曲线在A处为转折点，前后可分成二段直线，设一般的线性回归模型为：

$E(Y)=\beta_0+\beta_1x_1$ 现在我们用一个指示变量$x_2$，令$x_1<A$时$x_2=0$，$x_1>A$时 $x_2=1$，那么这一数据的分段线性回归方程可写成：

$$E(Y)=\beta_0+\beta_1x_1+\beta_2(x_1-A)x_2 \qquad (1)$$

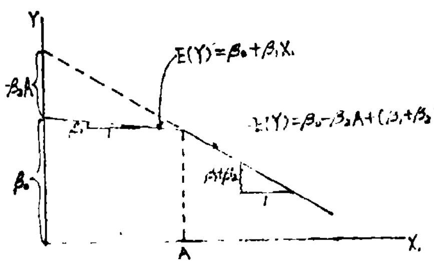
显然，当$x_1<A$时，把$x_2=0$代入公式(1)，上面的回归函数就成为：$E(Y)=\beta_0+\beta_1x_1$即以$\beta_0$为截距$\beta_1$为斜率的一条直线，当$x_1>A$时把$x_2=1$代入(1)式，其回归函数就成为：

$$E(Y)=\beta_0+\beta_1x_1+\beta_2(x_1-A)$$
$$=\beta_0-\beta_2A+(\beta_1+\beta_2)x_1$$

这是以$\beta_0-\beta_2A$为截距，$\beta_1+\beta_2$为斜率的一条直线。理论上三元线性回归是由因变量Y和自变量$X_1$、$x_2$组成的一个三维空间，为了直观起见，这是投影在Y和$x_1$平面上的二条直线，见图(3)

2、二个以上转折点的情况。上面的分段回归方法也可以推广到二个转折点以上，只要增加指示

图(3)一个转折点的分段线性回归




变量即可，现在假设有A和C二个转折点，这时就把曲线分割成三段直线，这时的回归模型可写成：

$$E(Y)=\beta_0+\beta_1x_1+\beta_2(x_1-A)x_2+\beta_3(x_1-C)x_3 \qquad (2)$$

定义指示变量$x_2$和$x_3$为：

若$x_1>A$时，$x_2=1$

若$x_1<A$时，$x_2=0$

又：若$x_1>C$时 $x_3=1$

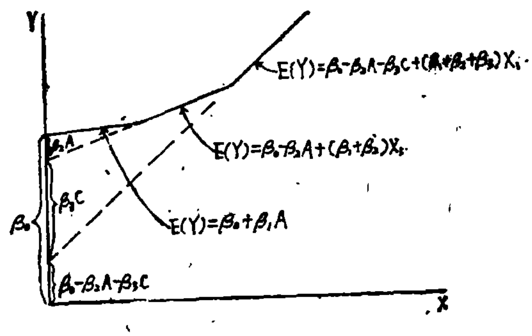
若$X_1<C$时 $x_3=0$

很显然，若$x<A$时，把$x_2=0$及$x_3=0$代入方程(2)得$E(Y)=\beta_0+x\beta_1$；若$C>x_1>A$时，把$x_2=1$和$x_3=0$代入方程(2)得：$E(Y)=\beta_0-\beta_2A+(\beta_1+\beta_2)x_1$；若$x>C$时，把$X_2=1$和$X_3=1$代入方程(2)得：

$$E(Y)=(\beta_0-\beta_2A-\beta_3C)+(\beta_1+\beta_2+\beta_3)x_1,$$

这是三条不同截距和斜率的三条直线，分别代表各线段的情况，可用图示如下：

图(4)二个转折点的分段线性回归




3、回归函数为不连续的情况。即如图(2)中所示那样，分段直线在D处有一个跳跃点。在这种情况下，可以用二个指示变量，设D处为跳跃点，则线性回归方程可写成：

$$E(Y)=\beta_0+\beta_1x_1+\beta_2(x_1-D)x_2+\beta_3x_3 \qquad (3)$$

定义指示变量$x_2$，$x_3$为：若$x_1>D$，则$x_2=1$，$x_3=1$ 若$x_1<D$时，则$x_2=0$，$x_3=0$。显然，在$x_1<D$时，把$x_2=x_3=0$代入方程(3)得

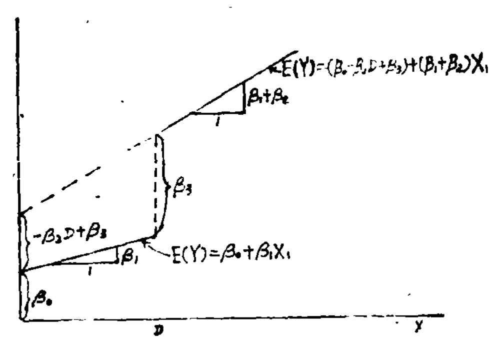
$E(Y)=\beta_0+\beta_1x_1$、

若$x_1>D$时，把$x_2=x_3=1$代入方程(3)得：

$$E(Y)=(\beta_0-\beta_2D+\beta_3)+(\beta_1+\beta_2)x_1,$$

该直线的截距为$\beta_0-\beta_2D+\beta_3$，斜率为$\beta_1+\beta_2$，这时的图形可以表示如下：

图(5)有一个不连续点的分段线性回归



## 三、分段线性回归的一个例子

现在我们用一个具体的例子来说明分段线性回归的用法。设有一工厂利用回归方法研究产品的单位成本(Y)与生产的批量(x)之间的关系，一般说来，生产的批量增加，则单位产品的成本下降，另外该厂有一台比较高级的新机器，当批量超过500件时可以使用，使成本有较大的降低，但若批量低于500件时，使用这机器并不合算，因而生产批量在超过500件使用新机器时有一转折点，从取得的数据也说明这一点：

根据下表资料可用一个转折点的分段回归，其转折点为$x_1=500$，令$x_2$为一指示变量，当$x_1>500$时，$x_2=1$，$x<500$时，$x_2=0$，其分段回归函数应为：

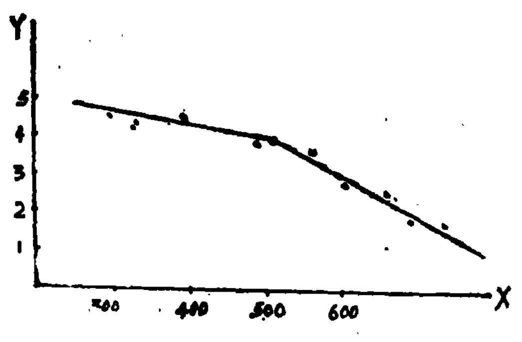
$$E(Y)=\beta_0+\beta_1x_1+\beta_2(x_1-500)x_2$$

令$x'_2=(x_1-500)x_2$ 则上式可写成：

$$E(Y)=\beta_0+\beta_1x_1+\beta_2x'_2$$

| 批号 i | 单位产品成本（元）$y_i$ | 生产批量（件）$x_i$ |
|---|---|---|
| 1 | 2.57 | 650 |
| 2 | 4.40 | 340 |
| 3 | 4.52 | 400 |
| 4 | 1.39 | 800 |
| 5 | 4.75 | 300 |
| 6 | 3.55 | 570 |
| 7 | 2.49 | 720 |
| 8 | 3.77 | 480 |

图（6）生产批量与单位成本之间的散点图




以上是一个典型的三元线性回归方程形式，我们可以用解三元回归方程方法求正规方程：

$$\begin{cases}\Sigma Y=nb_0+b_1\Sigma x_1+b_2\Sigma x'_2\\ \Sigma x_1Y=b_0\Sigma x_1+b_1\Sigma x_1^2+b_2\Sigma x_1x_2'\\ \Sigma x'_2Y=b_0\Sigma x_2'+b_1\Sigma x_1x'_2+b_2\Sigma x_2'^2\end{cases}$$

计算表

| $Y_i$ | $X_1$ | $X'_2=(X_1-500)X_2$ | $YX_1$ | $YX'_2$ | $\Sigma X_1^2$ | $\Sigma X'^2_2$ | $\Sigma X_1X'_2$ |
|---|---|---|---|---|---|---|---|
| 2.57 | 650 | 150 | 1670.5 | 385.5 | 422500 | 22500 | 97500 |
| 4.4 | 340 | 0 | 1496 | 0 | 115600 | 0 | 0 |
| 4.52 | 400 | 0 | 1808 | 0 | 160000 | 0 | 0 |
| 1.39 | 800 | 300 | 1112 | 417 | 640000 | 90000 | 24000 |
| 4.75 | 300 | 0 | 1425 | 0 | 900000 | 0 | 0 |
| 3.55 | 570 | 70 | 2023.5 | 248.5 | 324900 | 4900 | 39900 |
| 2.49 | 720 | 220 | 1792.8 | 547.8 | 518400 | 48400 | 158400 |
| 3.77 | 480 | 0 | 1809.6 | 0 | 230400 | 0 | 0 |
| 27.44 | 4260 | 740 | 13137.4 | 1598.8 | 2501800 | 165800 | 535800 |

故正规方程为

$$\begin{cases}8b_0+4260b_1+740b_2=27.44\\ 4260b_0+2501800b_1+535800b_2=13137.4\\ 740b_0+535800b_1+165800b_2=1598.8\end{cases}$$

解方程得$b_0=5.89545$　$b_1=0.00395$

$b_2=-0.00389$

故所求得分段线性回归方程为：

$$Y=5.89545-0.00395x_1-0.00389(x_1-500)x_2$$

假设我们要分别预计生产批量为440以及660件时的单位产品成本。当生产批量为440件时，$x_2=0$，单位产品成本预计为：

$Y=5.89545-0.00395(440)=4.15745$元，当生产批量为660件时，$x_2=1$，因此预计的单位成本为

$Y=5.89545+0.00389(500)-(0.00395+0.00389)660=2.66605$元。如果分段的数目比较多时，线性回归的变量也随着增多，当然用手工计算就比较困难了，但是正如前面已经提到，有大量的现成统计软件可以使用，也容易取得计算结果。

（作者单位　中国人民大学统计研究室）