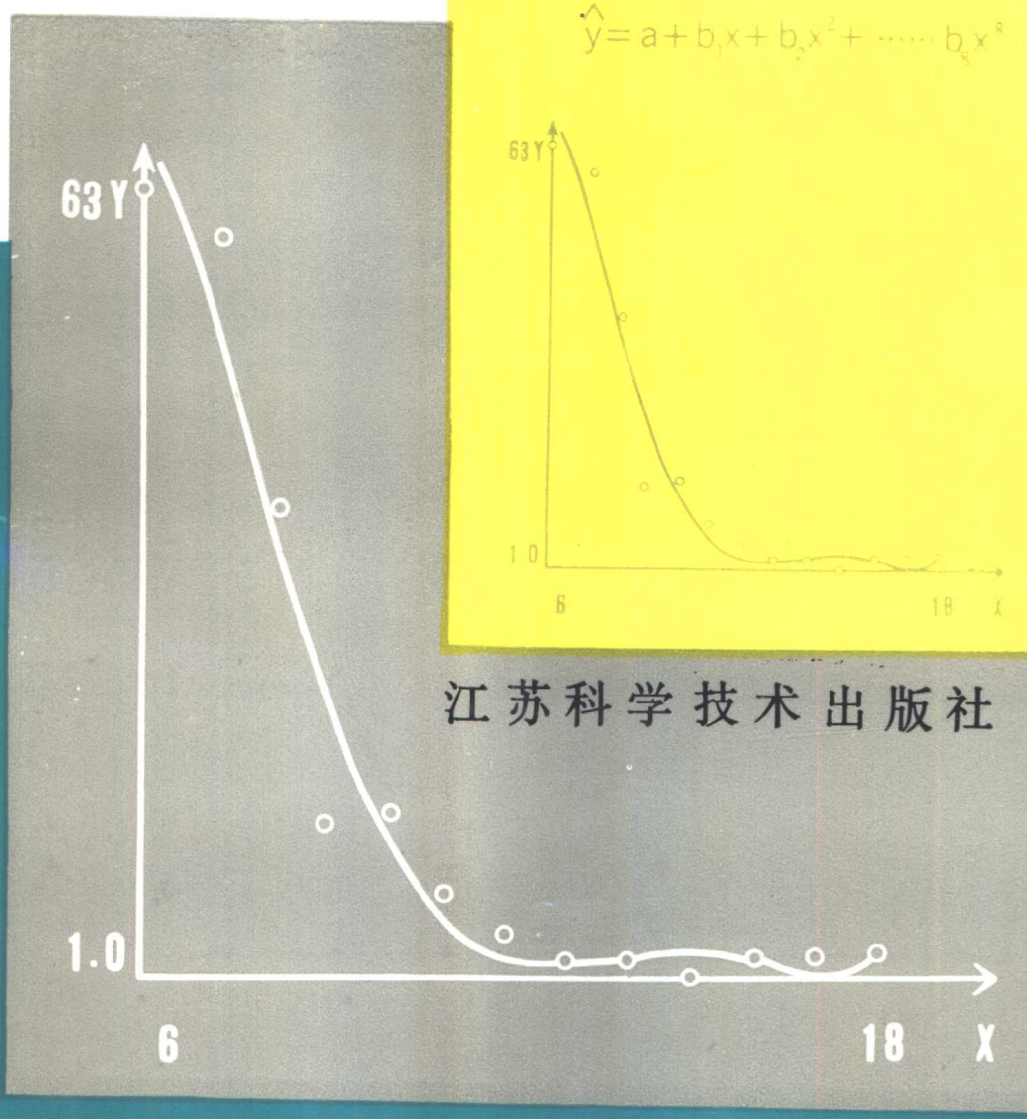
朱松成　崔　庶
许　鹰　崔宝善　编著

曲线回归优选



$$\hat{y}=a+b_1x+b_2x^2+\cdots\cdots b_sx^s$$

江苏科学技术出版社

曲线回归优选

李松成 崔 庶 许 鹰 崔宝善 编著

江苏科学技术出版社

曲线回归优选

朱松成　崔庶　许鹰　崔宝善　编著

出版发行：江苏科学技术出版社

印　　刷：常州人民印刷厂

开本787×1092毫米 1/16　印张7　插页2　字数167,000

1990年12月第1版　1990年12月第1次印刷

印数1—2000册

ISBN　7—5345—1065—1

R·165　　定价：3.90元

责任编辑　王　烈

序

曲线回归是科学实验和生产实践中研究事物内在规律的重要手段，它使人们由一般的定性分析上升到定量分析阶段。

但目前医疗系统在进行基础科学和临床医学实验时，所用的曲线回归方法还是采用五十年代手工描点作图，凭经验处理各种回归数据的原始方法，从而产生的误差也较大。朱松成等作者大胆提出了目前最新的计算优选回归法，使原先无法想象的庞大计算量的计算方法及数学模型逐步应用于实践之中，多种数学模型的优选也由理论的探讨变成实际应用。该方法在定量分析的基础上，彻底改变了过去工序繁杂、精度较低的手工回归法，自动从60多种函数模型中优选出最佳函数模型中优选出最佳函数模型进行回归计算。

该书用通俗的语言，以概念介绍与例题计算相结合的方法，给广大读者深入浅出地介绍了多种曲线回归方法、特点以及曲线拟合优度的具体指标，同时对数学模型的优选也作出了详细的介绍，并提出了电子计算机上实现曲线优选回归的具体程序，且使用方便、操作简单，产生的回归分析报告详细全面、曲线图型清楚直观。因此，本书具有较强的实用价值，也是我愿为之写序的原因。

南京军区后勤部卫生部部长 李富荣

一九九〇年十一月十五日

目　录

概　论

在客观世界中，变量与变量之间如果存在着一定的相关关系，那么这种关系可以分为确定性的关系和非确定性的关系两类。前者是当自变量x的值给定后，应变量y只有一个值与之对应，例如$y=\pi r^2$，式中r为半径，y为圆面积，这就是数学上的函数关系，不属于数理统计讨论的内容；后者是当自变量x确定之后，应变量y的值不是唯一确定的值，例如药物剂量与其反应之间，人的身高与体重之间，酶活性与pH值之间，细菌培养时间与细菌数量之间等，它们既存在着密切关系，又不能由一个变量的值精确地求出另一个变量的值，这就是数理统计学讨论的相关分析和回归分析。

19世纪英国遗传学家高尔登(Francis Galton)，在血缘关系的研究中发现：一组高个子父亲的儿子，比其父更高的概率要小于比其父矮的概率；同样，一组矮个子父亲的儿子，比其父更矮的概率亦小于比其父高的概率。这两组高矮不同父亲的儿子，在身高方面有着向总体“回归”的趋势。从此以后，人们便将回归一词作为统计研究事物相互关系的专用语。

定义　回归分析是指具有相互联系的事物或现象，根据其关系形式选用一个合适的数学模式，近似地表达其变量间平均变化关系的一种统计分析方法。

因此，回归分析是研究自变量与应变量之间的关系，求出回归方程式。相关分析是用相关系数(r)来度量回归方程式所描述的各个变量之间的密切程度。

如果直接用原始测定值，任意地将其连接，作回归分析，这是不可能正确反映两变量之间的关系，而往往得到错误的结论。也就是说，原始测定值是“实践”，必须上升为“理性认识”(拟合的曲线)，然后再用拟合的曲线即标准曲线指导工作(随机预测)，这是辩证唯物论的基本知识。

类型　回归分析是研究两个或两个以上变量间的关系形式，如果应变量与自变量的关系是线性的，则称线性回归分析；反之，称非线性回归分析。在线性回归分析中，可分为单元线性回归分析和多元线性回归分析。

一　单元线性回归分析　研究一个自变量与应变量相互关系，其模式为：

$$y=a+bx$$

式中x为自变量，y为应变量，a为截距，b为回归系数。

二　多元线性回归分析　研究多个自变量与应变量相互关系，其模式为：

$$y=a+b_1x_1+b_2x_2+\cdots\cdots+b_nx_n$$

式中b_1，b_2，……b_n，分别为y对x_1，x_2，……x_n的偏回归系数。

在单元线性回归中，当各实测点(x，y)密集在一条直线周围时，回归分析的任务就是采用一定的方法(常用最小二乘法)，求出一条实测点(变量值)，与该直线的纵向距离的平方和$[\Sigma(y-\hat{y})^2]$为最小值的直线方程，也就是配直线或直线拟合。如果自变量与应变量之间的分布不是直线关系(即非线性关系)时，就要选择恰当类型的曲线，使之更符合实际情况，所以称为配曲线，也叫曲线拟合。

用途　回归分析在医学研究、医院管理和日常工作中用途很广，常用于预测和控制。

1．寻找应变量(y)和自变量(x)之间的关系，并用回归方程来表达这种关系。例如，儿

科要按体重确定给药量，对1～7岁儿童的年龄与体重关系的回归方程式为：

$$y = 8 + 2x$$

2．确定两变量间的函数关系之后，由已知的变量值推估未知的变量值。例如，检验科根据回归方程从已知的吸光度来推估未知的谷丙转氨酶的单位。

3．对总体回归直线作出估计，说明样本回归直线的可信程度。

4．确定正常值范围。

第一章　回归分析

由于变量的离散性和随机性，在拟合直线或曲线时必须根据实测点的不同分布选用不同的数学模式，方能得出满意的回归结果，但是在实际工作中往往因计算工作量太大，而难以达到择优目的，因此，同一组标准样品实验数据，用不同方法建立标准曲线，其计算结果是不同的，在现有条件下，怎样选择适当的工具和方法建立标准曲线？用什么指标来衡量它的质量？这是一个带全局性的问题。为此我们提供了常用的直线、指数曲线、对数曲线、幂函数曲线、双曲线、S型曲线、Logistic曲线，以及多项式曲线、多项式对数曲线等60多种函数模型拟合标准曲线，采用人们容易接受的BASICA语言，按照模块式的结构，编制了计算机自动优选回归软件，采用汉字菜单提示的人机对话方式，从上述常用的多种数学模式中优选出拟合度最佳，剩余标准差($S_{y.x}$)最小的高质量标准曲线，而且自动地将各回归计算和假设检验的结果以及曲线的形状，实测点在回归曲线两侧分布状况和各预测值($\hat{y}$)报告出来，以此来探讨这一个带全局性的问题。现分别介绍如下：

一、直 线 回 归

在研究两个变量之间关系时，其散点图呈直线趋势，可作直线回归分析。

(一)直线回归分析的前提

1．两变量之间的关系要有实际意义。

2．一个变量是选定的，另一个变量是随机变量，且服从正态分布方可作回归分析。若两个变量都是随机变量，且服从双变量正态分布，则既可作回归分析，又可作相关分析。如果变量(一个或两个)呈偏态分布时，须作变量变换，使资料符合要求后再进行回归分析。

3．回归系数相当于解析几何中所说的斜率。用同一资料，由 x 推算 y：回归系数公式为 $b_{y.x}=\iota_{xy}/\iota_{xx}$，截距公式为 $a_1=\overline{y}-b_{y.x}\overline{x}$，回归方程为 $\hat{y}=a_1+b_{y.x}\overline{x}$；由 y 推算 x：回归系数公式为 $b_{x.y}=\iota_{x.y}/\iota_{yy}$，截距公式为 $a_2=\overline{x}-b_{x.y}\overline{y}$，回归方程为 $\hat{x}=a_2+b_{x.y}\overline{y}$。因此，要正确选定自变量，若两变量之间有因果关系，应以自变量如药物剂量为 x，反应率为应变量 y；无因果关系，应以易测者或变异较小者如仪器上读数为 x，药物浓度为应变量 y，否则误差增大。

4．回归方程只适用于原数据 x 所选择的范围之内，不得任意"外推"，而且工作条件也应保持与原数据所测定的条件一致。

5．变量值必须是同时、同地、同等条件下测得的，如果将两批数据混在一起作直线(曲线)回归是错误的。

例1.1　测得某地3岁儿童10人的体重与体表面积(表1.1)，试作相关分析和回归分析(数据取自《中国医学百科全书医学统计学》)。

表1.1　某地3岁儿童10人的体重与体表面积

体重 (Kg) X_i	11.0	11.8	12.0	12.3	13.1	13.7	14.4	14.9	15.2	16.0
体表面积(10^3cm^2) Y_i	5.283	5.299	5.358	5.292	5.602	6.014	5.830	6.102	6.075	6.411

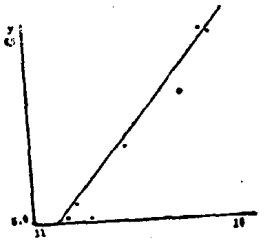


图1.1　某地8岁儿童体重与体表面积的散点图

1．将变量x、y分别作正态性检验(矩法)

H_0：总体服从正态分布，

H_1：总体非正态分布。

$\alpha = 0.05$。

$$G = \Sigma(x - \bar{x})^2/(n-1) \tag{1.1}$$

$$g_1 = [n\Sigma(x - \bar{x})^3]/[(n-1)(n-2)G^{3/2}] \tag{1.2}$$

$$g_2 = [n(n+1)\Sigma(x - \bar{x})^4 - 3(n-1)^3G^2]/[(n-1)(n-2)(n-3)G^2] \tag{1.3}$$

$$Sg_1 = \sqrt{6n(n-1)/(n-2)(n+1)(n+3)} \tag{1.4}$$

$$Sg_2 = \sqrt{24n(n-1)^2/(n-2)(n+5)(n^2-9)} \tag{1.5}$$

$$t_1 = g_1/Sg_1 \tag{1.6}$$

$$t_2 = g_2/Sg_2 \tag{1.7}$$

本例计算结果如下：

(1) 变量x的计算结果

$G = 24.904/(10-1) = 2.767111111$

$g_1 = (10 \times 2.79888)/(8 \times 9 \times 2.767111111^{3/2}) = 0.084452378$

$g_2 = [10 \times 11 \times 106.6238752 - 3 \times (10-1)^3 \times 2.767111111^2]/(9 \times 8 \times 7 \times 2.767111111^2)$

$= -1.30005678$

$Sg_1 = \sqrt{(6 \times 10 \times 9)/(8 \times 11 \times 13)} = 0.687042918$

$Sg_2 = \sqrt{[24 \times 10 \times 9^2]/[8 \times 15 \times (10^2-9)]} = 1.33424877$

$t_1 = 0.084452378/0.687042918 = 0.122921546$

$t_2 = -1.30005678/1.33424877 = -0.9743736$

(2) 变量y的计算结果

$G = 1.5438924/(10-1) = 0.1715436$

$g_1 = [10 \times 0.141036827]/[9 \times 8 \times (0.1715436)^{3/2}] = 0.275700944$

$g_2 = [10 \times 11 \times 0.387461225 - 3 \times (10-1)^3 \times (0.1715436)^2]/[9 \times 8 \times 7$

$\times (0.1715436)^2] = -1.46558618$

Sg_1与Sg_2同上。

$t_1 = 0.275700944/0.687042918 = 0.401286348$

$t_2 = -1.46558618/1.33424877 = -1.09843547$

(3) 判断结果

$u_{0.05} = 1.96$，x、y两变量的t_1与t_2皆小于$u_{0.05}$，$p > 0.05$，按$\alpha = 0.05$水准接受H_0、

拒绝H_1，故认为在偏度与峰度检验中总体是服从正态分布的。

2．求离均差平方和及离均差积和

(1) 求x的离均差平方和

$$l_{xx}=\sum(x-\overline{x})^2=\sum x^2-(\sum x)^2/n \tag{1.8}$$

$$=1831.24-(134.4)^2/10=24.904$$

(2) 求y的离均差平方和

$$l_{yy}=\sum(y-\overline{y})^2=\sum y^2-(\sum y)^2/n \tag{1.9}$$

$$=329.483368-(57.266)^2/10=1.5438924$$

(3) 求离均差积和

$$l_{xy}=\sum(x-\overline{x})(y-\overline{y})=\sum xy-(\sum x)(\sum y)/n \tag{1.10}$$

$$=775.5946-(134.4)(57.266)/10=5.93956$$

3．求相关系数(r)、回归系数(b)、截距(a)及列回归方程

(1) 求相关系数

$$r=\sum(x-\overline{x})(y-\overline{y})/\sqrt{\sum(x-\overline{x})^2\sum(y-\overline{y})^2}=l_{xy}/\sqrt{l_{xx}l_{yy}} \tag{1.11}$$

$$=5.93956/\sqrt{24.904\times1.5438924}=0.95788022$$

(2) 求回归系数

$$b=l_{xy}/l_{xx} \tag{1.12}$$

$$=5.93956/24.904=0.238498233$$

(3) 求截距

$$a=\overline{y}-b\overline{x} \tag{1.13}$$

$$=5.7266-0.238498233\times13.44=2.521183746$$

(4) 列回归方程

$$\hat{y}=a+bx \tag{1.14}$$

$$\hat{y}=2.5212+0.2385x$$

4．求估计误差平方和$[\sum(y-\hat{y})^2]$、剩余标准差$(S_{y\cdot x})$

(1) 求估计误差平方和

$$\sum(y-\hat{y})^2=l_{yy}-l^2_{xy}/l_{xx} \tag{1.15}$$

$$=1.5438924-(5.93956)^2/24.904=0.127317833$$

(2) 求剩余标准差

$$S_{y\cdot x}=\sqrt{\sum(y-\hat{y})^2/(n-2)} \tag{1.16}$$

$$=\sqrt{0.127317833/(10-2)}=0.126153593$$

5．作相关系数或回归系数的显著性检验

(1) 作相关系数显著性检验

$H_0:\rho=0$，

$H_1:\rho\neq0$。

$\alpha=0.05$。

$$t=r\sqrt{n-2}/\sqrt{1-r^2} \tag{1.17}$$

$$=0.9579\sqrt{10-2}/\sqrt{1-0.9579^2}=9.4369$$

查 t 界值表$t_{0.001(8)}=5.041$，$t>t_{0.001(8)}$，$P<0.001$。

按 $\alpha=0.05$ 水准拒绝 H_0、接受 H_1，认为该地3岁儿童体重与体表面积间有正相关。

或以 $r=0.9579$、$\nu=8$ 直接查相关系数界值表，$r_{0.001(8)}=0.872$，$r>r_{0.001(8)}$，$P<0.001$，结论相同。

*因为同一样本的相关系数与回归系数的显著性检验结果是相同的，故本例不必再作回归系数的显著性检验。如果两变量中只有一个是随机变量时，那么就必须作回归系数的显著性检验。

(2) 作回归系数显著性检验

$H_0:\beta=0$，

$H_1:\beta\neq0$。

$\alpha=0.05$。

$$SS_{回归}=\sum(\hat{y}-\bar{y})^2=bl_{xy}=l^2_{xy}/l_{xx} \qquad (1.18)$$

$$=5.93956^2/24.904=1.41657$$

$$SS_{剩余}=\sum(y-\hat{y})^2=l_{yy}-l^2_{xy}/l_{xx} \qquad (1.19)$$

$$=1.5438924-5.93956^2/24.904=0.12732$$

$$F=(SS_{回归}/1)/[SS_{剩余}/(n-2)] \quad \nu_1=1 \quad \nu_2=n-2 \qquad (1.20)$$

$$=(1.41657/1)/[0.12732/(10-2)]=89.0085$$

查F界值表，$F_{0.001(1,8)}=11.26$，$F>F_{0.001(1,8)}$，$P<0.001$。按 $\alpha=0.05$ 水准拒绝 H_0、接受 H_1，认为两变量之间存在着直线关系，可以求直线回归。

本书未作正态性检验的详细介绍，必要时请参阅有关医学统计学。

6. 计算器步骤

INV KAC INV Min MODE 2

11 x_D,y_D 5.283 DATA

11.8 x_D,y_D 5.299 DATA

12 x_D,y_D 5.358 DATA

12.3 x_D,y_D 5.292 DATA

13.1 x_D,y_D 5.602 DATA

13.7 x_D,y_D 6.014 DATA

14.4 x_D,y_D 5.83 DATA

14.9 x_D,y_D 6.102 DATA

15.2 x_D,y_D 6.075 DATA

16 x_D,y_D 6.411 DATA

Kout Σx^2 − Kout Σx INV x^2 + Kout n = 24.904(l_{xx})

Kout Σy^2 − Kout Σy INV y^2 + Kout n = 1.5438924(l_{yy})

Kout Σxy − Kout Σx × Kout Σy ÷ Kout n = 5.93956(l_{xy})

INV r 0.95788022 (r)

INV A 2.521183746(a)

INV B 0.238498233(b)

Kout Σy^2 − INV A × Kout Σy − INV B × Kout Σxy = 0.12731786 [$\Sigma(y-\bar{y})^2$]

0.12731786 ÷ 8 = INV $\sqrt{\ }$ 0.126153606 ($S_{y\cdot x}$)

11　　$\hat{y}$　5.145

11.1　$\hat{y}$　5.169

……

16　　$\hat{y}$　6.337

*凡是黑体的数据为计算器计算结果。

7. 直线与曲线拟合优度的比较　本例经计算机优选回归软件处理，得到7次抛物线为最佳拟合曲线，现将直线与7次抛物线(图1．2)进行比较，作方差分析如下：

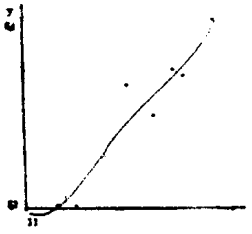


图1.2　实测点及拟合的7次抛物线图

H_0：曲线方次递升后减少的估计误差之均方等于升高曲线的剩余标准差，

H_1：曲线方次递升后减少的估计误差之均方大于升高曲线的剩余标准差。

$\alpha = 0.05$。

表1.2　方差分析表

拟合方式	$\Sigma(y-\hat{y})^2$	v	MS	F
直　　线	0.127	8		
7次抛物线	0.084	2	0.042	
差	0.043	6	0.007	6.17

F值＜1，P＞0.05，按照$\alpha = 0.05$水准接受H_0、拒绝H_1，认为直线与7次抛物线之间差异无显著性意义，但后者$\Sigma(y-\hat{y})^2$是前者的1/1.5，故也可优选。

二、非线性回归

在上一节中介绍了线性函数$y = a + bx$（a、b为常数），其散点图形呈直线分布，称为线性回归，所以在内容和计算中都比较简单方便。但在医学的实际问题中，x、y两变量之间的回归关系并非都是直线的，有的散点图呈曲线关系，或在一定范围内呈曲线关系，这时就应该选择适当类型的曲线进行拟合，故称为非线性回归。有一些非线性回归可以采用变量代换的方法，将比较复杂的函数化为线性函数，这就是所谓把曲线直线化的回归问题，或称为函数的线性化问题。因此，应该同样认真注意上一节“直线回归分析的前提”中所提出的有关问题。常用的有指数曲线、对数曲线、幂函数曲线、双曲线、S型曲线和logistic曲线等，现分别介绍如下：

(一)指数曲线

在单对数座标纸上作出的散点图，当y值取对数时，具有明显直线化趋势，便可用指数

函数进行拟合，并用拟合的指数曲线方程来分析两个变量之间的关系。其特点是在函数式转为直线式时肯定要将 y 值取对数，而不必对 x 值取对数，其渐近线与 y 轴平行，且曲线群以 y 轴的平行线为对称。指数曲线又称指数生长曲线，因为当自变量 x 值逐渐增大时，应变量 y 值便随之增大（或减少）得更快(图2.1)。

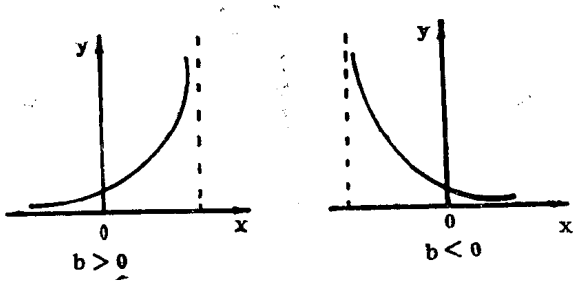


图2.1　指数曲线

指数曲线方程

1. $y = a e^{bx}$　　(2.1)

(2.1)公式两边各取以 e 为底的对数，有

$$\ln y = \ln a + b x$$

令　$\ln y = y'$　　$\ln a = a'$

则有 $y' = a' + b x$

2. $y = a e^{b/x}$　　(2.2)

(2.2)公式两边各取以 e 为底的对数，有

$$\ln y = \ln a + b/x$$

令　$\ln y = y'$　　$\ln a = a'$　　$1/x = x'$

则有 $y' = a' + b x'$

3. $y = e^{a+bx}$　　(2.3)

(2.3)公式两边各取以 e 为底的对数，有

$$\ln y = a + b x$$

令　$\ln y = y'$

则有 $y' = a + b x$

例2.1　某防治站重复治疗钩虫病人的次数与复查阳性率资料(表2.1)，试拟合指数曲线(数据取自《医学统计方法》)。

表2.1　钩虫治疗次数与复查阳性率关系

治疗次数	x_i	1	2	3	4	5	6	7	8
复查阳性率(%)	y_i	63.9	36.0	17.1	10.5	7.3	4.5	2.8	1.7

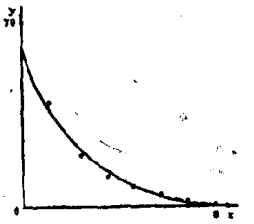


图2.2　钩虫治疗次数与复查阳性率的散点图

解：

1．作散点图(图2.2)。

2．a、b 值及指数曲线方程

(1) 求 a、b 值

根据(1.8)、(1.9)、(1.10)、(2.1)公式求得

$\overline{x}=4.5$

$\overline{y'}=2.247936429$

$\Sigma(x-\overline{x})(y'-\overline{y'})=-21.2624197$

$\Sigma(x-\overline{x})^2=42$

$\Sigma(y'-\overline{y'})^2=10.84407459$

根据(1.12)公式求得

$b=-21.2624197/42=-0.50624808$

根据(1.13)公式得

$a'=2.247936429-(-0.50624808)\times 4.5=4.526052829$

取ln a′的反对数

$e^{a'}=e^{4.526052829}=92.39314881$

(2) 计算器步骤

MODE 2　INV KAC　INV Kin

1　x_D,y_D　63.9　ln　DATA

2　x_D,y_D　36.0　ln　DATA

3　x_D,y_D　17.1　ln　DATA

4　x_D,y_D　10.5　ln　DATA

5　x_D,y_D　7.3　ln　DATA

6　x_D,y_D　4.5　ln　DATA

7　x_D,y_D　2.8　ln　DATA

8　x_D,y_D　1.7　ln　DATA

INV $\overline{x}$　4.5　($\overline{x}$)

INV $\overline{y}$　2.247936429　($\overline{y'}$)

Kout Σxy − Kout $\Sigma x\times$ Kout $\Sigma y\div$ Kout n = −21.2624197($l_{xy'}$)

Kout Σx^2 − Kout Σx INV $x^2\div$ Kout n = 42(l_{xx})

Kout Σy^2 − Kout Σy INV $x^2\div$ Kout n = 10.844077459($l_{y'y'}$)

INV B　−0.50624808　(b)

INV A INV e^x 92.39314882 （a）

(3) 回归方程

根据(2.1)公式得

$$\hat{y} = 92.39314882e^{-0.50824808x}$$

3. 求剩余标准差

(1) 求估计误差平方和

$$\begin{aligned}\Sigma(y-\hat{y})^2 &= (63.9-55.69023069)^2+(36-33.56744341)^2\\ &\quad+(17.1-20.23287107)^2+(10.5-12.19542002)^2\\ &\quad+(7.3-7.350823769)^2+(4.5-4.430729732)^2\\ &\quad+(2.8-2.670634826)^2+(1.7-1.609732663)^2\\ &=86.03923882\end{aligned}$$

(2) 计算器步骤

1 $\hat{y}$ INV e^x 55.69023069 − 63.9 = INV x^2 M+

2 $\hat{y}$ INV e^x 33.56744341 − 36 = INV x^2 M+

3 $\hat{y}$ INV e^x 20.23287107 − 17.1 = INV x^2 M+

4 $\hat{y}$ INV e^x 12.19542002 − 10.5 = INV x^2 M+

5 $\hat{y}$ INV e^x 7.350823769 − 7.3 = INV x^2 M+

6 $\hat{y}$ INV e^x 4.430729732 − 4.5 = INV x^2 M+

7 $\hat{y}$ INV e^x 2.670634826 − 2.8 = INV x^2 M+

8 $\hat{y}$ INV e^x 1.609732663 − 1.7 = INV x^2 M+

MR 86.03923882

(3) 求剩余标准差

根据(1.16)公式求得

$$S_{y\cdot x} = \sqrt{86.03923882/(8-2)} = 3.786802495$$

4. 曲线的拟合优度

(1) 求相关指数

$$\begin{aligned}R^2 &= 1-\Sigma(y-\hat{y})^2/\Sigma(y-\bar{y})^2 \qquad (2.4)\\ &= 1-86.03923882/3281.335 = 0.973779196\end{aligned}$$

相关指数接近1，表明指数曲线拟合度甚佳。

(2) 方程的拟合效果检验

本例经计算机优选回归软件处理，得6次抛物线（图2.3）为最佳拟合曲线，现将指数曲线与6次抛物线进行比较，作方差分析如下：

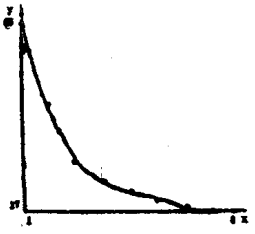


图2.3 实测点及拟合的6次抛物线图

H_0：曲线方次递升后减少的估计误差之均方等于升高曲线的剩余方差。

H_1：曲线方次递升后减少的估计误差之均方大于升高曲线的剩余方差。

$\alpha = 0.05$。

表2.2 方差分析表

拟合方式	$\Sigma(y-\hat{y})^2$	v	MS	F
指数曲线	86.04	6		
6次抛物线	2.00	1	2.00	
差	84.04	5	16.81	8.41

查F界值表，$F_{0.05(5,1)} = 230$，$P > 0.05$，按$\alpha = 0.05$水准接受H_0、拒绝H_1，认为指数曲线与6次抛物线之间差异无显著性意义。但是后者的$\Sigma(y-\hat{y})^2$只有前者的1/43，故可优选。

例2.2 某单位在某地作卫生调查时，得儿童各年龄组麻疹曾患率（%）资料见表2.3，试拟合指数曲线(数据取自《医学统计方法》)。

表2.3 各年龄儿童组与麻疹曾患率关系

年　龄	x_i	1	2	3	4	5	6
曾患率(%)	y_i	34.3	65.5	76.8	85.2	90.3	94.1

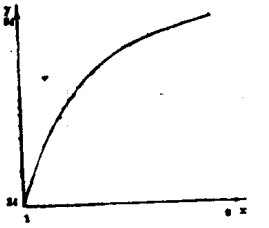


图2.4 各年龄儿童组与麻疹曾患率的散点图

解：

1. 作散点图(图2.4)。

2. 求a、b值及指数曲线方程

根据(1.8)、(1.9)、(1.10)、(2.2)公式求得

$\overline{x'} = 0.408333333$

$\overline{y'} = 4.258482846$

$\Sigma(x' - \overline{x'})(y' - \overline{y'}) = -0.59077347$

$\Sigma(x' - \overline{x'})^2 = 0.490972222$

$\Sigma(y' - \overline{y'})^2 = 0.71227169$

根据(1.12)公式求得

$b = -0.59077347/0.490972222 = -1.20327269$

根据(1.13)公式求得

$a' = 4.258482846 - (-1.20327269)(0.408333333) = 4.7498192$

取$e^{a'} = e^{4.7498192} = 115.5633888$

(2) 计算器步骤

MODE 2 INV KAC INV Min

1 INV 1/x x_D, y_D 34.3 ln DATA

2 INV 1/x x_D, y_D 65.5 ln DATA

3 INV 1/x x_D, y_D 76.8 ln DATA

4 INV 1/x x_D, y_D 85.2 ln DATA

5 INV 1/x x_D, y_D 90.3 ln DATA

6 INV 1/x x_D, y_D 94.1 ln DATA

INV $\overline{x}$ **0.408333333**($\overline{x}'$)

INV $\overline{y}$ **4.258482846**($\overline{y}'$)

Kout Σxy − Kout Σx × Kout Σy ÷ Kout n = **-0.59077347**($l_{x'y'}$)

Kout Σx^2 − Kout Σx INV x^2 ÷ Kout n = **0.490972222**($l_{x'x'}$)

Kout Σy^2 − Kout Σy INV x^2 ÷ Kout n = **0.71227169**($l_{y'y'}$)

INV A **4.7498192** INV e^x **115.5633888** (a)

INV B **-1.2032727** (b)

(3) 回归方程

根据(2.2)公式得

$\hat{y} = 115.5633888e^{-1.2032787/x}$

3. 求剩余标准差

(1) 求估计误差平方和

$$\Sigma(y - \hat{y})^2 = (34.3 - 34.69329688)^2 + (65.5 - 63.31883573)^2 + (76.8 - 77.37999609)^2 + (85.2 - 85.54144747)^2 + (90.3 - 90.84589954)^2 + (94.1 - 94.56370645)^2 = 5.87817182$$

(2) 计算器步骤

1 INV 1/x $\hat{y}$ INV e^x **34.69329688** − 34.3 = INV x^2 M+

2 INV 1/x $\hat{y}$ INV e^x **63.31883573** − 65.5 = INV x^2 M+

3 INV 1/x $\hat{y}$ INV e^x **77.37999609** − 76.8 = INV x^2 M+

4 INV 1/x $\hat{y}$ INV e^x **85.54144747** − 85.2 = INV x^2 M+

5 INV 1/x $\hat{y}$ INV e^x **90.84589954** − 90.3 = INV x^2 M+

6 INV 1/x $\hat{y}$ INV e^x **94.56370645** − 94.1 = INV x^2 M+

MR **5.87817182**

(3) 求剩余标准差

根据(1.16)公式求得

$S_{y \cdot x} = \sqrt{5.87817182/(6-2)} = 1.212247068$

4. 曲线拟合优度

(1) 求相关指数

根据(2.4)公式求得