徐崇民 著

化曲线为直线的回归分析问题

（一元曲线回归分析）

厦门大学出版社

化曲线为直线的回归分析问题

（一元曲线回归分析）

徐崇民

厦门大学出版社

图书在版编目(CIP)数据

化曲线为直线的回归分析问题/徐崇民著. —厦门:厦门大学出版社,1999.9
ISBN 7-5615-1531-6

Ⅰ.化…　Ⅱ.徐…　Ⅲ.回归分析-研究　Ⅳ.O212.1

中国版本图书馆 CIP 数据核字(1999)第 02735 号

厦门大学出版社出版发行
(地址:厦门大学　邮编:361005)
厦门新嘉莹彩色印刷有限公司印刷
(地址:厦门市莲前北路 77 号　邮编:361009)
1999 年 9 月第 1 版　1999 年 9 月第 1 次印刷
开本:850×1168　1/32　印张:3　字数:63 千字
定价:8.00 元

序

作者1944年毕业于福建农业大学前身——省立农学院，获农学士学位。毕业后留校任教，讲授《生物统计》、《田间实验设计与分析》、《统计学》等课程，课余时间进行多种田间实验，4年间在省级学术刊物发表论文3篇。1948年任讲师，1961年历任教务处副处长、系副主任。

1976年退休，在保健锻炼之余，大量阅读有关科技书刊，积累资料，仔细运算编写，先后完成《田间实验方法》、《多元线性回归分析》、《一元曲线回归分析》等诸著初稿，约10万多字，这些稿件曾提供给有关单位作为教学科研之用。有关人士曾建议投稿专业刊物或作为生物统计教学参考书籍出版，惟因经费及印刷等原因而搁置多年。嗣后继续阅读近年有关科技书刊，对一元曲线回归问题做进一步分析和研究，写成《化曲线为直线的一元回归分析问题》。本书涉及曲线类型较多，分析比较深入，在工农业生产和科研中应用广泛，因而具有实用意义。经征求专家意见，提请厦门大学数学系教授审阅，经进一步修改后，出版作为生物统计学的参考用书。限于本人数理水平，谬误之处，敬请指正。

本书先后经过福建农业大学郭如灼教授、厦门大学数学系陈传淡教授审阅，并提供宝贵修改意见，又承蒙厦门市计划委员会经济研究所戴松若博士进行数据验算，最后经厦门大学出版社精心编印，一并致谢。

徐崇民

1999年5月于福建农业大学

目　录

一

线性化的意义和步骤

化曲线为直线的回归分析问题，是回归分析中的一个重要方法问题。

回归分析是应用数理统计方法，对处于一定条件的客观事物所反映的数量性质进行大量观测，分析研究它们之间相互依存的统计关系，建立相应的回归方程（或叫经验公式），用来比较确切地说明问题和预测预报或控制该事物有关性质可能出现的取值范围，因此在工农业生产和科学研究中常常广泛应用它，属性性质往往可以代换为数量进行探讨。

在回归分析中，根据自变量的个数和函数的图像，分有一元线性回归（又叫简单回归）、多元线性回归（又叫多重回归）；一元曲线回归和多元回归。化曲线为直线回归分析问题，是指在方法上可以把变量间的非线性关系问题，通过一定的变量代换化成线性关系，来计算有关参数估计量，然后建立曲线回归方程，这种方法简称函数的线性化。

一般地说，化曲线为直线的回归分析方法可分三个步骤进行。

（一）选定因变量与自变量之间内在关系的函数类型

在工农业生产和物理学、化学、生物学等自然科学中常常遇到的是幂函数曲线和指数函数曲线的回归关系。

选定的方法，一种是根据专业知识，从理论上推导或根据以往积累的实际经验，确定两个变量之间的函数关系类型。例如在一种称为生长现象的生物实验中，如在一定条件下某种细菌培养，每一时刻的细菌总量（y）与某个阶段繁殖时间（x）的关系，根据专业知识判断是 $y=ae^{bx}$ 的指数关系，其中 y 是因变量，x 是自变量，e 是自然对数的底，a、b 为待定的未知参数估计量。另一种是根据理论或经验还无法推导的两个变量之间的函数类型，先用实验数据绘制散点图，然后从散点图的分布形状和特点选择适当的函数类型来拟合，必要时可多选一二种作比较。

（二）确定相关函数中未知参数的估计量，建立曲线回归方程

在往往可以拟合的幂函数和指数函数中，对未知参数的估计，一般可以应用线性化方法将呈现为曲线关系的实验数据，进行倒数变换或对数变换等化为直线关系的数据，然后用确定线性方程未知参数的最小二乘法或平均值法等来确定变换成线性方程的未知参数，最后再根据原来变换的关系反过来算出拟合函数中未知参数的估计量，建立曲线回归方程。

如对上述指数函数 $y=ae^{bx}$，取自然对数得

$$\ln y=\ln a+bx$$

命 $\ln y=y'$，$\ln a=a'$，变换成线性方程

$$y'=a'+bx$$

其中 a' 是直线在 y 轴上的截距，b 是回归系数。然后用最小二乘法确定 a'、b'，最后反过来算出 a 并整理出指数函数曲线回归方程

$$\hat{y}=ae^{bx}$$

其中 a、b 是确定了的实数。

线性化的具体方法，将在以下探讨。

（三）检验拟合回归方程的效果，如有需要，作出必要的预测预报或控制

回归分析的目的不仅是建立回归方程，以利于比较确切地说明因变量和自变量之间的依存关系，更重要的是应用已经建立的回归方程，根据自变量的取值范围，预测预报因变量的估计值；或根据因变量的要求来控制自变量，亦即通过对自变量的控制以达到控制因变量的目的。但在建立回归方程之后，要对拟合的回归方程进行效果检验和比较，选用拟合效果较好的回归方程，那么根据它作出的预报或控制就有意义。在预测预报或控制中采用内插法为宜，外推法可能出现较大的偏差，不宜使用。

下面将对十多种简单而比较广泛应用的曲线函数的图像特点、线性化法和回归分析方法进行讨论，说明图像特点的目的是提供选择拟配曲线函数数学模式的参考。叙述回归分析方法，主要是结合示例，对有的数学模式比较系统地讨论了几种不同方法，应用时可根据资料情况，

选用其一;有时利用同一资料是为了选择较好的数学模式。为了减少篇幅,对部分示例略去检验或预测预报,应用时应根据需要,参考其他示例的检验和预测预报方法进行。

二

幂函数的图像特点和回归分析

设自变量为 x,因变量为 y,用自变量 x 的某一个幂来表示的函数,其形式如 $y=x^k$ 的叫幂函数。幂的指数 k 可以是正负整数、零或正负分数,随着 k 值的不同,函数的图像也不同。根据幂函数的图像,可将它分为两大类,一是双曲线函数,一是抛物线函数。

(一)双曲线函数

图像在渐近线对角中表现为相互对称的两条曲线的函数叫双曲线函数。如当 $|y|=|x|^k$ 的 $k=-1$ 时,$|y|=\frac{1}{|x|}$。x 和 y 成反比关系,它的图像见图 2.1。可以看到,当 x 的绝对值开始增大时,y 的绝对值逐渐减少;当 x 的绝对值增大到一定程度时,y 的绝对值减小的速度逐渐缓慢;当 $|x|\to 0$ 时,$|y|\to\infty$;当 $|x|\to\infty$时,$|y|\to 0$。因此它是以 x 轴和 y 轴为渐近线的双曲线。当 x 取正值时,图像在第一象限,开口朝右上方。因为在农业和生物学上,x 多数取正值,因此本文主要讨论第一象限的图像。

双曲线函数几种比较常见的形式、图像特点、线性化法及其回归分析。

$$y=a+b\frac{1}{x} \tag{2.1}$$

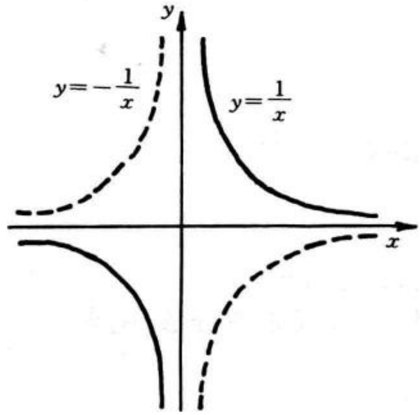

图 2.1　$y=\pm\frac{1}{x}$的图像

图像见图 2.2，水平渐近线为直线 $y=a$，竖直渐近线为 y 轴，在 $0\leqslant x<+\infty$ 区间，当 $a>0$、$b>0$ 时，曲线的开口朝右上方，y 值与 x 值成反比型；当 $a>0$、$b<0$ 时，曲线的开口朝右下方，y 值与 x 值成正比型。

线性化法是对(2.1)或令 $x'=\frac{1}{x}$，得线性方程

$$y=a+bx'$$

用最小二乘法等方法确定式中 a 和 b。

例 1　陕西土壤肥料研究所对土壤硝化力(y)与速效

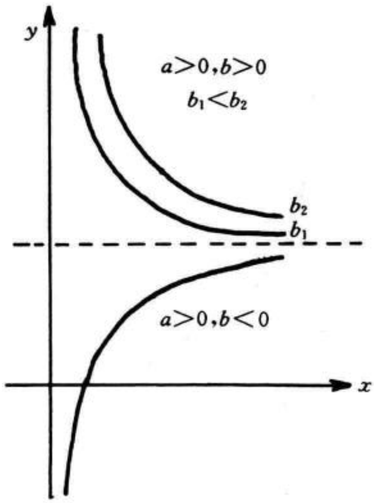

图 2.2　$y=a+b\dfrac{1}{x}$的图像

磷含量(x)的关系,采用 $y=\dfrac{x}{a+bx}$进行分析。① 本人就其散点图估测部分散点的近似值作为观测值(见表 2.1)来说明拟合 $y=a+b\dfrac{1}{x}$的回归分析法。

第一,绘制散点图,选择拟配曲线函数。

散点图见图 2.3。先拟配 $y=a+b\dfrac{1}{x}$。

第二,确定未知参数估计量,建立回归方程。

根据线性化列表计算,方法有二。

① 《中国农业科学》,(4):52,1978

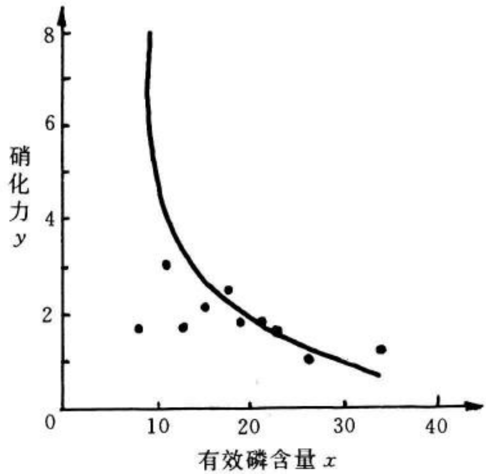

图 2.3 土壤硝化力与有效磷含量关系

(1)最小二乘法　方法是使估计值与实际值之间的差距为最小，$\sum(y-\hat{y})^2=min$。以此为基础构建联立方程。

$$\begin{cases} y=a+bx' \\ \sum y=na+b\sum x'^2 \end{cases}$$

求出 a 和 b 值。

$$b=\frac{\sum x'y-(\sum x')(\sum y)/n}{\sum x'^2-(\sum x')^2/n} \qquad (2.2)$$

$$=\frac{1.735\ 896-0.616\ 036\times 23.1/10}{0.043\ 704\ 1-0.616\ 036^2/10}$$

$$=54.368\ 49$$

$$a=(\sum \bar{y})/n-b(\sum x')/n$$

$$= \bar{y} - b\,\bar{x}' \tag{2.3}$$

$$= 23.1/10 - (54.368\ 49)(0.616\ 036)/10$$

$$= -1.039\ 29$$

表 2.1　回归系数计算表

x	y	$\frac{1}{x}=x'$	$x'y$	x'^2
9	6.6	0.111 111	0.733 333	0.012 345 6
11	3.0	0.090 909	0.272 727	0.008 264 1
13	1.7	0.076 923	0.130 769	0.005 917 1
15	2.1	0.066 667	0.140 000	0.004 444 4
17	2.3	0.058 824	0.135 294	0.003 460 1
19	1.8	0.052 632	0.094 737	0.002 770 0
21	1.8	0.047 619	0.085 714	0.002 268
23	1.6	0.043 478	0.069 565	0.001 890 3
26	1.0	0.038 462	0.038 462	0.001 479 2
34	1.2	0.029 412	0.035 294	0.000 865 0
188	23.1	0.616 036	1.735 896	0.043 704

将 a、b 数值代入(2.1)式，得双曲线回归方程

$$\hat{y} = 54.368\ 49\,\frac{1}{x} - 1.039\ 29$$

水平渐近线为直线 $y=a=-1.039\ 29$；竖直渐近线为 y 轴。

(2)平均值法(亦叫分割和法)　方法是将观测资料按自变量由小到大顺序排列后，将资料按未知参数的个

数平分两组，如不能平分，则第一组可多分一行（见表2.2）。应用下式计算未知数估计量。

表 2.2　分割和表

行数	x	$x'=\frac{1}{x}$	y
1	9	0.111 111	6.6
2	11	0.090 909	3.0
3	13	0.076 923	1.7
4	15	0.066 667	2.1
5	17	0.058 823	2.3
小计	65	0.404 433	15.7
6	19	0.052 631	1.8
7	21	0.047 619	1.8
8	23	0.043 478	1.6
9	26	0.038 461	1.0
10	34	0.029 412	1.2
小计	123	0.211 601	17.4
总计	188	0.616 034	23.1

$$\begin{cases}\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i \\ \sum_{i=n+1}^{k} y_i = (k-n)a + b\sum_{i=n+1}^{k} x_i\end{cases} \tag{2.4}$$

n 为第一组观测个数，$k-n$ 为第二组观测个数。在本例为

$$\begin{cases}15.7=5a+0.404\ 433b\\7.4=5a+0.211\ 601b\end{cases}$$

两个方程相减并解之得

$$b=43.042\ 6$$

在计算第二个未知参数估计量时，为了避免计算误差，应以联立方程的综合式

$$\sum_{i=1}^{nk} y_i=(n+k)a+b\sum_{i=1}^{nk} x_i \tag{2.5}$$

在本例将表 2.2 底数值代入(2.5)式，为

$$23.1=10a+0.616\ 034b$$

$$a=\frac{23.1-0.616\ 034(43.042\ 6)}{10}$$

$$=-0.341\ 6$$

代入(2.1)式，得双曲线回归方程

$$\hat{y}=43.043\ 6\frac{1}{x}-0.341\ 6$$

平均值法计算比较简单，但对观测值误差较大的资料，拟合误差亦较大，宜慎用。

第三，检验拟合效果，并作出必要的预测或控制。检验方法有几种，现分述如下，应用时可选其一。

(1)方差分析法　对线性化的线性回归方程进行方差分析，用 F 值检验，如 F 值显著，说明这种线性化是适宜的，相应表示拟配的曲线回归方程是有效的。如F值不显著，说明这种拟合不适宜，可能这两个变量之间的关系不是曲线关系或不是这种函数类型，或关系不明显，应另行分析。本例计算见表 2.3，并应用下列各式计算各种平