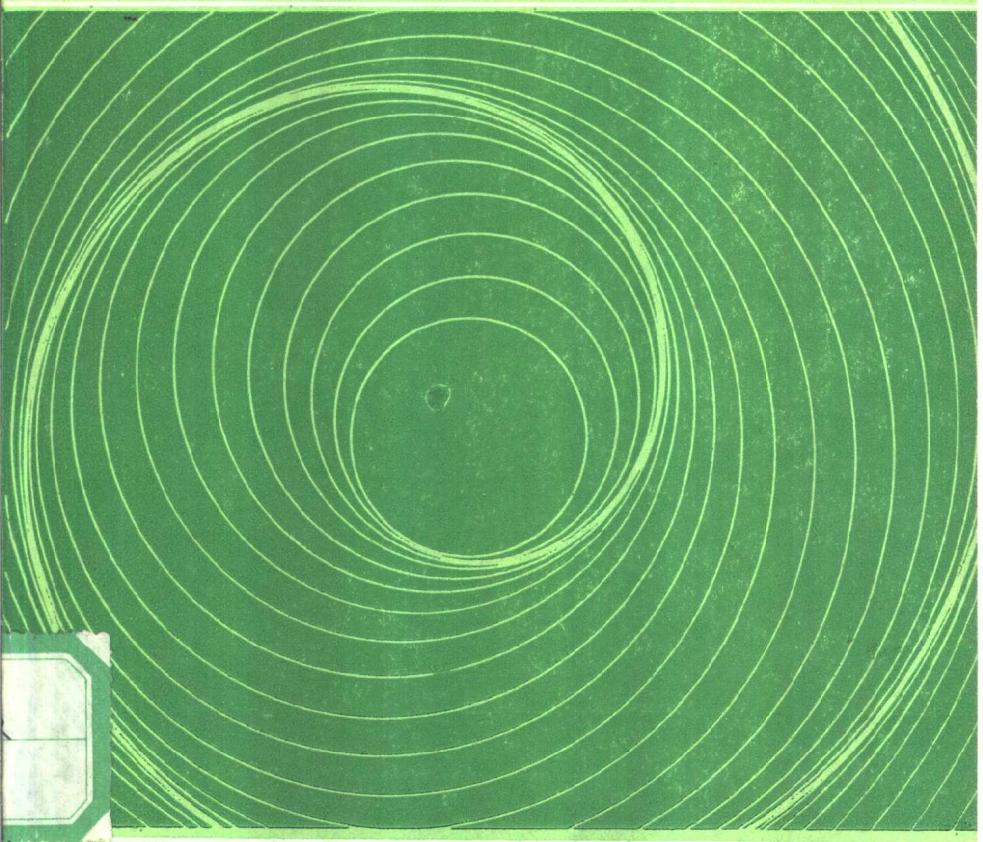
正交多项式回归及其应用

薛 为 民 编著



化学工业出版社

正交多项式回归及其应用

薛 为 民 编著

化学工业出版社

内 容 提 要

正交多项式回归是数理统计的一个重要方法。

本书作者从多项式回归的基本概念出发，结合实例，对多项式回归方法作了较全面的介绍。作者还结合其应用正交多项式回归方法的经验，论述了计算实践中常遇到的一些问题及解决方法。书末附有完整的多项式回归计算表可供查阅。

本书可作为学习多项式回归方法的入门书，也可供实际工作者和教学工作者参考。

正交多项式回归及其应用

薛为民编著

责任编辑：邓原

封面设计：任辉

化学工业出版社出版发行

（北京和平里七区十六号楼）

化学工业出版社印刷厂印刷

新华书店北京发行所经销

开本787×1092 1/32 印张8 1/4 字数178千字

1989年11月第1版 1989年11月北京第1次印刷

印 数 1—2,300

ISBN 7-5025-0482-6/O·7

定 价4.70元

序

近几年来，薛为民同志在推行企业现代化管理过程中，联系了化工、医药、丝绸、机械等行业的实际，在大量实践的基础上，设计出了新正交多项式计算表。与此同时，他还提出了“插值法”、“平均值法”和“分组中值法”扩展了原正交多项式回归方法只适用于自变量取等间距的使用范围，使原来繁琐的计算得到了简化。

在本书中，作者介绍了他的成果及设计原理，并收集了大量的应用实例。内容涉及一元线性回归，多元线性回归和非线性回归。文字简洁，实例丰富。本书可以作为初学回归分析的入门书，也可作为一般从事数据处理实际工作及教学工作者的参考书。

华东工学院 管理工程系
教授 吴云敏
1987年1月

前　言

当今，运用数理统计的手段已日益成为现代化管理的需要和基本要求。而正交多项式回归则是数理统计的一个重要方法。它在非线性模型的建立中，占着重要的地位。它充分利用正交函数的特点，有效地克服了古典回归方法上的弱点，使其在回归分析上具有特殊的作用。目前国内通用的正交多项式计算表是从国外引进的。实践表明它存在着运算繁难和易出差错的弊病，因而阻碍了它的推广运用。为减少大量的计算量，在实践基础上，我对原表及算法作了修正和改进，充分利用函数 $\Phi_i(x)$的对称和反对称原理，简化了计算过程，革除了繁琐费时的高次多项式的整理化简过程，减少了差错，节约了时间，提高了效率。多年来概率统计上一直认为，正交多项式表系只有在自变量取等间隔的条件下才能适用，而通常情况下这种条件只有在实验时自变量严格控制下才有可能得到满足。但从生产现场直接获得或在试验中随机获得的自变量由于不是等间隔，因此不能应用这方法而限制了其优越性的发挥。为了克服上述缺陷，扩展正交多项式表系的适用范围，近年来，我结合化工等行业的生产实际，在实践中对其进行了探索和研究，设计出了‘新正交多项式计算表”及计算程序，并提出了“插值法”、“平均值法”和“分组中值法”三种方法，使自变量为不等间隔的数据，不但能运用此表系计算，而且还克服了传统方法的缺点使回归方程具有较高的精度和稳定性。这就是我编写本书的目的。

本书按章、节体例编写，力求做到章节结构合理，阐述计算准确。全书共分六章，主述正交多项式回归，也涉及到一元线性回归、多元线性回归和非线性回归、多项式回归。书末附有作者独创的新正交多项式计算表等数理统计用表，内容较为丰富，易于自学，可供化工等行业科技人员阅读，也可作为从事数据处理工作者和教学工作者的工具参考书。特别是书中结合应用实例简略地介绍了我国质量管理专家张公绪教授的最新研究成果——选控控制图原理与方法，所以，十分适合于推行全面质量管理过程中广大质量管理工作者和质量教育培训的参考书。

本书在编写过程中，自始至终得到化工部生产综合司和质量处领导和有关同志的大力支持和关心，得到化工部杨馨洁副总工程师，江苏省化工质量管理协会秘书长范映星，华东工学院吴云从教授，北京钢铁学院张公绪教授、中国人民大学沈思聪教授，苏州大学陈庆云副教授的热忱指导，此外还得到朱玉君、薛苏生、薛太存、石应津、朱朝阳等同志的帮助，在此一并表示衷心的感谢！化工出版社的同志曾给予大力协助与指导，亦顺致谢意！

因本人水平有限，疏漏之处在所难免，敬请各界读者批评指正。

作者于1987年1月

目　录

第一章　多项式回归

第一节　回归分析方法简述

回归分析是一种重要的数理统计方法，它有着十分广泛的用途，受到人们越来越高度的重视。在工农业生产以及科研中许多问题都可以用这个方法解决，生物学、化学、新产品研制开发时试验数据的处理，经验公式的求得，因素分析，产品质量控制，某些新标准的制订，气象、地震预报，自动控制中数学模型的建立等许多场合，回归分析都得到广泛的应用。甚至在经济分析和经济预测中，回归分析也成了必不可少的手段。而化工生产是一个多因素，多变量，多指标的复杂系统，为了寻求最优化的生产条件，达到优质高产，低消耗的目的，回归分析则是十分有效的工具之一。

比如：在某化工产品的生产过程中，原料T在催化剂的作用下，经过一段时间的蒸汽加热转化为产品t，为了提高T的转化率，必须寻求时间、温度和催化剂浓度之间最适当的组合，而在此之前就必须设法建立转化率对于时间、温度和催化剂浓度之间的数学模型，这就可以借助于回归分析方法。具体地讲，回归分析的主要任务就是通过试验或观测数据来寻找诸变量之间的统计关系，并且运用这种统计关系，从一个或几个变量所取的值去有效地预测与它们相关的另一个随机变量的值，它包括：①从一组数据出发确定这些变量之间的定量关系式，这种关系式也可以称做回归函数；②对关系式的可信程度进行检验；③从影响着某一个量的许多变

量中判断哪些变量的影响是显著的，哪些是不显著的（这问题对于一元回归是不存在的）；④利用所求得的关系式对某一过程进行预报和控制。

在回归分析中，根据自变量的多少，有一元回归和多元回归之分。仅有一个自变量的称为一元回归，自变量在两个或两个以上的称为多元回归。

从回归分析所研究的数学模型看，主要是线性回归模型或多项式回归模型，后者又可转化为前者，但后者本身又有一些特殊的方法。它的主要任务是解决非线性模型的建立问题，它们都是基于数学上的“最小二乘法原理”。本书主要是论述多项式回归模型的建立中的正交多项式回归问题，但叙述中也必然地涉及到了一元和多元线性回归的计算问题。

严格地讲，当因变量与自变量均为随机变量时叫做相关分析，而其中有一个为非随机变量时叫回归分析，但通常将二者不加区别统称为回归分析，因为它们在数学处理的方法上是一致的。

生物学家F. Galton在遗传学的研究中首先引入了“回归”的概念。他于1889年在《普用回归定律》一文中说：“每个人的特点是和他的亲属相共的，但平均地说在程度上要差一点”。他的朋友K. Pearson收集了1078个家庭中儿子身高与父亲身高的数据，作出了散点图，由图看出，虽然个子高的父亲确有生高个子儿子的趋向，但是一群高个子父亲的儿子们的平均高度都低于父亲们的平均高度，而一群矮个子的父亲的儿子们的平均高度都高于父亲们的平均高度，这就是说，高个子或矮个子的儿子们的高度有一个回归，即回复于全体男子的平均高度的倾向。K. Pearson由资料求得的儿子身高（y, cm）对于父亲身高（x, cm）的直线回归方程

为：$\hat{y}=0.516x+84.33$，由于这个历史的原因，“回归”一词一直沿用至今。但其含义早已突破了遗传学的范畴。

第二节　多项式回归的基本概念

在实际问题中，两个变量之间的内在关系往往是非线性的，也就是曲线的关系，从广义上说，线性关系是非线性关系的特殊情况。所以在回归分析中，我们除了要研究线性回归外，还应对非线性回归的计算加以研究。通常，人们是通过某种线性化处理的方式将非线性回归转化为线性回归的问题，其中一种方法是根据因变量和自变量各个观察值作成的曲线图选取适当的函数类型进行拟合。所选函数中未知参数的确定，可以通过对自变量或因变量（或同时对二者）进行适当的变量变换，把曲线方程化为直线方程。这种方法确实能够解决一部分的非线性回归问题，但也存在一些问题。比如，究竟选择什么样的函数类型，对于回归方程的精度影响很大，实际计算中往往需要对多种函数类型的回归方程进行比较，才能获得较为满意的结果。其次，并不是所有的问题都是能够通过变量变换化曲线为直线的方法所能解决的。

例如　$y=a+bx+cx^2+dx^3+\cdots\cdots$　　(1-1)

就不能通过变量变换把它化为直线。但是，微积分的知识告诉我们，任何函数至少在一个比较小的邻域内可以用多项式逼近。所以在通常比较复杂的实际问题中，可以不问因变量与自变量的关系如何而用多项式回归进行分析计算。

设有一组观察值 (x_t, y_t)，$(t=1,\cdots\cdots,n)$，若它们是非线性关系，我们就可以用一个K次多项式来拟合

$$y=a_0+a_1x+a_2x^2+\cdots\cdots+a_Kx^K \qquad (1-2)$$

假定我们设定$\varphi_1(x)$，$\varphi_2(x)$，$\cdots\cdots$，$\varphi_K(x)$分别是x的一

次、二次及K次多项式，则（1-2）式也可用$\varphi_i(x)$来表示

$$y=d_0+d_1\varphi_1(x)+d_2\varphi_2(x)+\cdots\cdots+d_K\varphi_K(x) \qquad (1-3)$$

将$\varphi_i(x)$看作是新变量，（1-3）式就是一个K元线性回归方程其回归系数d_i由下面的正规方程确定

$$\begin{Bmatrix} l_{11}l_{12}\cdots\cdots l_{1K} \\ l_{21}l_{22}\cdots\cdots l_{2K} \\ \cdots\cdots\cdots \\ l_{K1}l_{K2}\cdots\cdots l_{KK} \end{Bmatrix} \begin{Bmatrix} d_1 \\ d_2 \\ \vdots \\ d_K \end{Bmatrix} = \begin{Bmatrix} l_{1y} \\ l_{2y} \\ \vdots \\ l_{Ky} \end{Bmatrix} \qquad (1-4)$$

又 $$d_0=\bar{y}-d_1\bar{\varphi}_1(x)-d_2\bar{\varphi}_2(x)\cdots\cdots d_K\bar{\varphi}_K(x) \qquad (1-5)$$

其中 $$l_{ij}=\sum_{t=1}^{N}[\varphi_i(x_t)-\bar{\varphi}_i(x)][\varphi_j(x_t)-\bar{\varphi}_j(x)]$$

$$i、j=1,2\cdots\cdots,K$$

$$L_{iy}=\sum_{t=1}^{N}[\varphi_i(x_t)-\bar{\varphi}_i(x)][y_t-\bar{y}]$$

这样就把多项式回归的问题转化成多元线性回归的问题计算了。同样，多元多项式回归问题也可以化为多元线性回归来解决，比如，对于一个包括多变量的任意多项式回归模型

$$y=d_0+d_1z_1+d_2z_2+d_3z_1^2+d_4z_1z_2+d_5z_2^2+\cdots\cdots \qquad (1-6)$$

只要令$x_1=z_1$，$x_2=z_2$，$x_3=z_1^2$，$x_4=z_1z_2$，$x_5=z_2^2\cdots\cdots$就可以将其化为多元线性回归问题来解决。

当然，多项式模型并不一定都是如（1-2）式所示的K次多项式模型，其中也可以包括诸如$\sqrt{x_i}$，$\lg x_i$等项，但它们也仍然可以化为线性回归的问题。

第三节　多项式回归的计算

多项式回归的计算，通常可以分成以下几个步骤：

1．确定多项式的表达形式

通常多项回归可选用如（1-2）式所示的一个K次多项式来拟合，即$y=a_0+a_1x+a_2x^2+\cdots\cdots+a_Kx^K$，$K$大小的选取可根据散点图中估计的曲线的形状与复杂程度和拟定的方程的精度而定。一般情况下，对于一个抛物线，起码选取二次多项式（即$K=2$），对于三次曲线起码选取三次多项式（即$K=3$），也就是K的值起码要大于曲线的峰（谷）数。K取得大拟合精度可以提高，但计算量也随之增大，而且过分地追求精度的提高，往往也没有什么实际价值，因为还必须考虑到试验误差的大小，经济的办法是，精度保证在误差范围内而K也尽可能地选得小一些。

如前所述多项式的表达形式除如（1-2）式所示外，其中还可以有诸如x^{-1}，x^{-2}，$x^{1/2}$，$x^{-1/2}$，$(x^2-1)^{1/2}$，$\ln x$，e^x，e^{-x}，$\sin x$，$\sin^2 x$，$\cos x$，$\cos 2x$等初等函数项。

2．求出多元线性回归模型的回归系数

通过变量变换，例如　$x=x_1$，$x^2=x_2$，$\cdots\cdots x^K=x_K$，或如（1-3）式，$\Phi_1(x)=x_1$，$\Phi_2(x)=x_2$，$\cdots\cdots$，$\Phi_K(x)=x_K$，则可获得广义的多元线性回归模型

$\hat{y}=d_0+d_1x_1+d_2x_2+\cdots\cdots+d_Kx_K$，其中回归系数$d_i$由求解如（1-4）式所示的正规方程而得。

3．回归方程效果检验

在多元线性回归中回归平方和表示的是所有K个自变量对y的变量的总影响，（多项式回归中，多项式的每一项视为一个自变量）。按下式计算

$$U=\Sigma(\hat{y}-\bar{y})^2=\sum_{i=1}^{K}d_i l_{iy}$$

而剩余平方和（各点残差平方和）则是除这些自变量外，其它随机因素对y的影响，按下式计算

$$Q=\Sigma(y-\hat{y})^2=l_{yy}-U$$

各平方和的自由度按下面的原则确定：总平方和l_{yy}的自由度为$N-1$，N为总试验次数，回归平方和的自由度等于自变量的个数K，从而剩余平方和的自由度等于$N-1-K$，剩余标准差按下式计算

$$s=\sqrt{\frac{Q}{N-1-K}}$$

利用下式计算回归均方与剩余均方的比，即统计量F值

$$F=\frac{U/K}{Q/(N-K-1)}$$

然后查F分布表对整个回归方程进行显著性检验。

上述情况，可以总结在一个方差分析表中。

方差分析表

变差来源	平方和	自由度	均方	F值
回归	$U=\sum_{i=1}^{K}d_i l_{iy}$	K	U/K	U/KS^2
剩余	$Q=l_{yy}-U$	$N-K-1$	$S^2=\frac{Q}{N-K-1}$	
总计	$l_{yy}=\Sigma(y-\bar{y})^2$			

4．检验每一个因子在多项式回归方程中的作用

我们知道，回归平方和是所有自变量对于变差的总贡献。

因此，若去掉一个自变量后回归平方和减少的数值越大，说明该因子在回归中起的作用越大，我们称取消一个自变量后回归平方和减小的数值为 y 对这个因子的偏回归平方和，记作 P_i，可用下式计算

$$P_i=\frac{d_i^2}{c_{ii}}$$

对各因子的分析可按两步进行：

（1）先计算统计量 F 值，对各因子的显著性进行检验。

$$F_i=\frac{P_i}{S^2}=\frac{d_i^2}{c_{ii}S^2}$$

其中 S^2 是方差分析中的剩余方差，F_i 的自由度为 $1, N-K-1$，在给定的显著水平 α 时，查 F 分布表就可以检验该因子的偏回归平方和的显著性了。

（2）经 F 检验后对那些不显著的因子从回归方程中剔除，此时回归系数 d_i 应重新计算，此时因子的偏回归平方和的大小也有所改变，故还应对它们重新检验。根据检验结果确定多项式回归模型的最终表达形式。现举例说明：

例1-1 已知某种半成品在生产过程中其废品率 y 与

表 1-1 废品率 y 与化学成分 x 的实测数据表

序号	1	2	3	4	5	6	7	8
x(0.01%)	34	36	37	38	39	39	39	40
y(1%)	1.30	1.00	0.73	0.90	0.81	0.70	0.60	0.50
序号	9	10	11	12	13	14	15	16
x	40	41	42	43	43	45	47	48
y	0.44	0.56	0.30	0.42	0.35	0.40	0.41	0.60

它的某种化学成分 x 有关，表1-1中列出了一一对应 的实 测数据试求 y 对 x 的回归方程。

从图1-1可以看出，废品率最初随着成分的增加而降低，而当成分超过一定值之后，又有所回升。根据这个特性，考虑抛物线

$$y=d_0+d_1x+d_2x^2$$

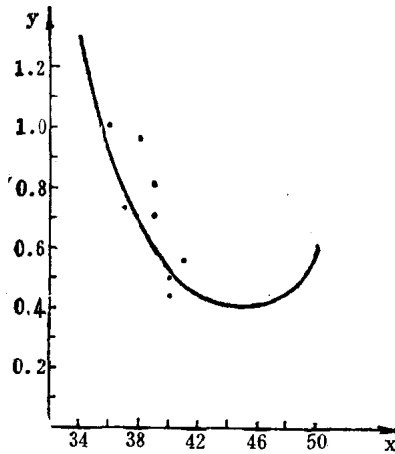


图 1-1　废品率与化学成分含量关系图

令 $x_1=x$，$x_2=x^2$，则有 $y=b_0+b_1x_1+b_2x_2$，这样多项 式回归问题就转化成二元线性回归的计算问题了。对每个 x 计算相应的 x^2 作为第二个自变量，然后进行计算，结果如下

$$\Sigma y=10.02 \quad \bar{y}=10.02/16=0.6263$$

$$\Sigma x_1=651 \quad \bar{x}_1=651/16=40.6875$$

$$\Sigma x_2=26709 \quad \bar{x}_2=26709/16=1669.3125$$

这些计算都是在一种规格化的表上进行的，便于检查，不易出错，见表1-2。

表 1-2　多项式回归计算表

序号	$x_1=x$	$x_2=x_1^2$	y	x_2^2	x_1y	x_2y	y^2
1	34	1156	1.30	1336336	44.2	1502.8	1.69
2	36	1296	1.00	1679616	36	1296	1.00
3	37	1369	0.73	1874161	27.01	999.37	0.5329
4	38	1444	0.90	2085136	34.2	1299.6	0.81
5	39	1521	0.81	2313441	31.59	1232.01	0.6561
6	39	1521	0.70	2313441	27.3	1064.7	0.49
7	39	1521	0.60	2313441	23.4	912.6	0.36
8	40	1600	0.50	2560000	20	800	0.25
9	41	1600	0.44	2560000	17.6	704	0.1936
10	42	1681	0.56	2825761	22.96	941.36	0.3136
11	42	1764	0.30	3111696	12.6	529.2	0.09
12	43	1849	0.42	3418801	18.06	776.85	0.1764
13	43	1849	0.35	3418801	15.05	647.15	0.1225
14	45	2025	0.40	4100625	18	810	0.16
15	47	2209	0.41	4879681	19.27	905.69	0.1681
16	48	2304	0.60	5308416	28.8	1382.4	0.36
Σ	651	26709	10.02	713370681	396.04	15803.46	7.3732
$\frac{1}{n}\Sigma$	40.6875	1669.3125	0.6263				

$$l_{11}=\sum x_1^2-\frac{1}{n}(\sum x_1)^2=221.44$$

$$l_{22}=\sum x_2^2-\frac{1}{n}(\sum x_2)^2=1513685$$

$$l_{12}=l_{21}=\sum x_1x_2-\frac{1}{n}(\sum x_1)(\sum x_2)=18283$$

$$l_{1y}=\sum x_1y-\frac{1}{n}(\sum x_1)(\sum y)=-11.649$$