病态模型的统计诊断

李俊 著

BING TAI MO XING DE TONG JI ZHEN DUAN

航空工业出版社

病态模型的统计诊断

李俊 著

航空工业出版社

北京

内 容 提 要

本书主要通过非约束与约束两种方法对病态模型进行研究，在非约束条件部分通过初等变换和投影变换两种方法介绍了病态模型的影响分析；在约束条件部分主要建立了病态模型数据删除前后相应模型岭估计及方差参数估计之间的关系。本书共十章，第一章和第二章分别对病态模型作了概述及介绍后面各章所用到的基础知识；第三章至第六章分别介绍了非约束病态模型的参数估计及非约束条件下病态模型的影响分析；第七及第八章介绍了约束条件下病态模型的影响分析；最后两章是关于协方差矩阵扰动生长曲线模型的影响分析。

本书可作为应用统计、管理科学及从事实际应用的大学生、研究生、教师、科技人员和统计工作者的参考书。

图书在版编目（CIP）数据

病态模型的统计诊断 / 李俊著. -- 北京 : 航空工业出版社，2013.11

ISBN 978-7-5165-0267-9

Ⅰ. ①病… Ⅱ. ①李… Ⅲ. ①数理统计－研究 Ⅳ. ①O212

中国版本图书馆 CIP 数据核字(2013)第 272169 号

病态模型的统计诊断

Bingtai Moxing de Tongji Zhenduan

航空工业出版社出版发行

（北京市朝阳区北苑路 2 号院 100012）

发行部电话：010-84936555 010-64978486

北京市科星印刷有限责任公司印刷 全国各地新华书店经售

2013 年 11 月第 1 版 2013 年 11 月第 1 次印刷

开本：787×1092 1/16 印张：15 字数：365 千字

印数：1－1000 定价：48.00 元

序

统计诊断是近年来迅速发展起来的一门统计学新分支。它以强烈的应用背景、新颖的统计思想、广泛的研究内容和丰富的实际成果呈现出一个理论与应用紧密结合的崭新领域。顾名思义，统计诊断就是对实际问题中得到的数据、提炼出的模型及相应统计推断方法的合理性进行研究，检查数据、模型及其推断方法中可能存在的“毛病”，并提出相应的“治疗”措施。这里的“毛病”主要是针对既定模型与数据集的拟合情况而言的，而“治疗”措施则是根据既定模型与数据集的拟合情况所作出的肯定或否定既定模型的对策。

文献［1］已结合非约束条件线性模型[①]较为详细地介绍了统计诊断的原理和方法，其内容包括：线性回归的异常点分析、线性回归的残差分析、线性回归的影响分析（或总体影响分析、数据点的影响分析）、数据变换及其诊断、广义影响分析基础（即局部影响分析）、回归诊断的 Bayes 方法等。

文献［2］就模型的诊断也介绍了诸如残差图诊断检验、异方差性诊断检验及拟合欠佳诊断检验等内容。在残差图诊断检验部分介绍了点图诊断和删除法、残差图误区及用残差图评估模型等；在异方差性诊断检验方面提供了参数回归中异方差性推断、非参数回归中异方差性诊断检验及半参数方差函数回归模型异方差性诊断检验等；在拟合欠佳诊断检验方面给出了同方差情形下拟合欠佳检验及异方差情形下拟合欠佳检验等内容。

另外，文献［2］在模型的局部影响分析部分也介绍了一些独到的见解，如关于线性模型系数的局部影响分析、关于变换的局部影响分析、关于残差平方和的局部影响分析、关于多重势的局部影响分析及纵向数据随机效应模型参数的局部影响分析等内容。

文献［3］则从异方差性、自相关性及多重共线性问题方面评估了模型的拟合状况，并且对异方差性、自相关性及多重共线性问题的产生、危害及如何识别和诊断等作了较为详细的介绍。

所有这些内容为评价模型的拟合状况提供了多姿多态的诊断方法和衡量标准，在不同程度上进一步丰富了统计诊断的理论。大量的研究和应用实践使人们对统计诊断的意义和价值有了明确的认识，目前，统计诊断已广泛应用于各种统计问题和许多统计模型，并被编入许多通用的软件包，成为统计学使用过程中一个不可缺少的重要步骤。

非病态线性模型统计诊断的有关理论已广泛应用于工程、农林、经济、医学、教育、心理及管理科学等领域，随着应用的不断深入，促进了这些领域的发展；同时，由于实践的不

---

① 非约束条件线性模型即设计阵为非退化的模型，本文称其为非约束条件线性模型，其型号是对设计阵趋于退化的模型而言的。设计阵趋于退化的模型又可称为带约束条件的线性模型或病态模型.

断深入，客观上要求理论本身进一步丰富和完善，这样又促进了统计诊断理论的研究。

随着实践的不断发展，人们又发现了一种新的模型——设计阵趋于退化的模型，也称为病态模型。它是人们从某些涉及定性变量问题中提炼出来的一种模型。生长曲线模型及方差分析模型就是这种模型的典型代表。

由于客观事物联系的普遍性，我们常常很难找到甚至根本不存在一组对同一因素都有影响的不同变量是无关的；而且由于认识上的局限性，对究竟有多少个变量在影响着我们所关心的问题不是太了解。为了获得尽可能多的信息，人们往往偏向于“宁多毋少”的原则来选择自变量，这样就会产生信息的重叠，这种信息的重叠现象在模型中的体现就是其设计阵趋于退化。

另外，由于事物的运动规律一般不可能在短时期内充分暴露出来，它需要有一个过程，因此，在一有限时期内收集到的数据就可能出现重复现象，当然更谈不上“足够”的数据了，这种情况在涉及多个自变量时尤为突出。这种数据的简单重复现象也将导致设计阵趋于退化。

由此看来，在实际中处理设计阵趋于退化的模型问题是不可避免的。这样现实就对我们提出了一个新的问题：如何建立有关设计阵趋于退化模型的统计诊断理论？

病态模型与非病态模型是基于其设计阵是否为退化进行区分的，从纯矩阵的角度而言，退化阵远比非退化阵要广泛的多，非退化阵只不过是退化阵中的特殊情形。从这个意义上讲，非病态模型是病态模型的特例，病态模型是非病态模型的推广。

不论是非病态模型，还是病态模型，它们首先都是模型。如果仅有非病态模型的理论，而缺乏病态模型的理论，那只能说明有关模型的理论还欠完善。不论是从理论上，还是实践上讲，病态模型的理论都应列入我们的研究范畴。对病态模型理论的研究可以使已建立的非病态模型的理论得到进一步的丰富和完善，同时，已有模型的发展和应用的加深也将促进对病态模型理论的研究。

文献［4］为了解决设计阵趋于退化的线性模型回归系数 β 的估计 $\hat{\beta}$ 性质将变差这一情形，通过对 β 各分量加以限制的方法即椭球约束获得了设计阵趋于退化的线性模型的岭估计，并进一步研究了该模型的一些性质，如有偏性及 MDE（mean dispersion error）有效性等。

文献［5］为了研究协方差矩阵发生扰动时生长曲线模型岭估计的影响分析，建立了生长曲线模型、协方差矩阵扰动生长曲线模型岭估计之间的关系，讨论了协方差扰动和数据删除对岭估计的影响等。为了回答例如工业上用 a 种不同工艺加工钢材时不同工艺的优劣等类似问题，文献［6］引入了带有线性约束条件 $h'\beta=0$ 的方差分析模型。由于方差分析模型的设计阵一般是呈退化的，因此，问题归结为研究带双约束条件 $h'\beta=0$ 和 $\beta'N\beta\leqslant 1$ 的线性模型。以上内容为我们研究病态模型有关理论打下了基础。

现在的问题是，我们如何建立以上这种病态模型的影响分析，以及其统计诊断理论？

另外，虽然文献［4］通过对病态模型设计阵的回归系数 β 各分量加以限制的方法，获得了其岭估计。但是，应该承认，这样的岭估计是不可靠的，甚至是值得怀疑的。因为这种岭估计实际是向原点压缩的有偏估计，它仅仅是为了考虑当模型设计阵 X 趋于退化，引起其回归系数 β 的估计 $\hat{\beta}$ 将在 q 维空间的某些方向上偏大这一不良性质而对其进行改进所获得的估计。既然是改进，那我们当然要问，这种改进的程度如何？究竟要改进到何种程度才使岭估计最佳？等等。

根据文献［1］，在模型的影响分析中，研究数据点的影响主要研究其回归系数 β 估计 $\hat{\beta}$ 的影响，甚至在有关模型异常点的检验部分，也通过 $\hat{\beta}_{(i)}-\hat{\beta}$ 是否有显著差异来衡量第 i 个数据点是否异常。当然我们完全可以研究数据点对其他参数估计的影响。但是，在回归分析中，与其他参数相比，回归系数 β 的估计 $\hat{\beta}$ 充当着更重要的角色，有不可替代的作用。

显然，如果我们执意要以病态模型的岭估计为基础去解决其他问题，那么所获得的结论就会与实际相差甚大，甚至导致错误的统计推断。如果这样的话，研究病态模型的有关理论首先要解决的问题就是如何去获取病态模型的最佳估计，否则，有关这一模型的其他研究就无从进行。可这样的最佳估计是否存在呢？如果存在，又以什么标准进行衡量呢？以上这些问题都是我们所要研究的内容。

在本书的研究中，文献［4］、［5］、［6］为本书提供了研究对象——模型，文献［1］、［2］、［3］则为我们研究此类模型的统计诊断理论提供了一定的方法。本书共十章，大致可分为四个部分：

第一部分（第一章和第二章）分别对病态模型及后面各章所涉及的知识进行了概述和介绍。

第二部分（第三章至第六章）通过矩阵的初等变换分解法获得了非约束病态模型的参数估计，从而为解决病态模型的影响分析打下了基础；通过综合相关变量的方法解决了新模型的应用问题；此外通过所获得的与病态模型等价的模型——P.T.等价模型与 E.T.等价模型，借助于其影响分析解释了病态模型的影响分析及病态模型异常点的检验问题。

第三部分（第七章及第八章）通过约束法解决了病态模型的影响分析，其中，解决了岭参数的估计问题，建立了病态模型数据删除前后相应参数（主要是回归系数及方差参数）估计之间的关系；获得了约束病态模型的广义似然函数及在各种不同扰动下对应模型的广义似然函数；在约束病态模型的局部影响分析部分获得了其在各种扰动下的影响曲率 C_{d}、最大影响曲率 $C_{\max}$ 及最大影响方向 $d_{\max}$ 的具体表达式。

第四部分（第九章和第十章）是关于协方差矩阵扰动生长曲线模型的影响分析。

本书是在作者的硕士毕业论文的基础上经过多次修改及补充写成的，主要分为三个阶段：第一阶段，该毕业论文于 2005 年初以四篇文章的形式投到了《应用概率统计》刊物上，

审稿老师提出了宝贵的修改意见和有待解决的遗留问题；第二阶段，着力解决已投文章中的遗留问题；第三阶段，在2009年初被派往上海师大进修，在进修期间，撰写了两篇关于如何确定岭参数的文章，同时解决了病态模型的广义似然函数问题。考虑到岭估计是一个向原点压缩的有偏估计，且变数较大，不宜于以此为基础对病态模型作进一步的研究，因此，2011年又通过非约束的方法获得了比岭估计更优良的非约束估计，从而为进一步研究病态模型打下了良好的基础。

由于本人水平有限，书中难免存在不妥之处，在此恳请同行专家及广大读者提出批评和建议。

作　者

2013年8月25日

目　录

第一章　引　论

众所周知，线性模型(linear model)是已获得深入研究且应用广泛的模型．而事实上，线性模型只不过是一个较理想化的模型，它的一些假定条件，如方差满足齐性及设计阵为非退化阵等条件，在实际中往往得不到满足，因此需要对其定义进行延拓，对其有关理论进行补充和完善，以便满足实践发展的需要．

1.1　非病态线性模型的基本理论

非病态线性模型是相对于病态模型而言的，它实际上就是我们通常所指的设计阵为非退化阵的线性模型．非退化阵即列满秩阵，退化阵即列降秩阵．为了更好地理解病态模型的产生、存在及探索研究病态模型的方法，在此，我们对线性模型的有关理论作简要介绍，有关该模型更详细的内容可参见文献[1]和[7].

1.1.1　非病态线性模型的一般形式

非病态线性模型的一般形式为

$$y_i=\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_{q-1}x_{i,q-1}+\varepsilon_i,i=1,2,\cdots,n, \tag{1.1.1}$$

其中，y_i 为因变量，$x_{ij}(1\leqslant j\leqslant q-1)$为自变量，$\varepsilon_i$ 为随机误差．模型(1.1.1)式的第 i 组观测值可记为$(y_i;x_{i1},x_{i2},\cdots,x_{i,q-1})$．模型(1.1.1)式也可用矩阵形式表示如下

$$Y=X\beta+\varepsilon, \tag{1.1.2}$$

其中，$Y=(y_1,y_2,\cdots,y_n)^T$ 为 $n\times 1$ 阶观测阵，X 为 $n\times q$ 阶列满秩设计阵，其第 i 行为$(1,x_{i1},x_{i2},\cdots,x_{i,q-1})$，$\beta=(\beta_0,\beta_1,\beta_2,\cdots,\beta_{q-1})^T$ 为 $q\times 1$ 阶未知回归系数阵，$\varepsilon=(\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n)^T$ 为 $n\times 1$ 阶随机误差阵．

对于模型(1.1.2)式的随机误差项 ε，一般假定其分量 $\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n$ 相互独立，数学期望为零，方差满足齐性，即

$$E(\varepsilon)=0,var(\varepsilon)=\sigma^2 I, \tag{1.1.3}$$

其中，σ^2 为未知常数，I 为 n 阶单位阵．(1.1.3)式通常记为如下形式

$$\varepsilon\sim(0,\sigma^2 I). \tag{1.1.4}$$

在大多数情况下，假定 ε 服从标准正态分布，则

$$\varepsilon\sim N(0,\sigma^2 I). \tag{1.1.5}$$

1.1.2 非病态线性模型的参数估计及其性质

模型(1.1.2)式在满足条件(1.1.4)式下的回归系数 β 的最小二乘估计为

$$\hat{\beta}=(X^TX)^{-1}X^TY. \tag{1.1.6}$$

方差参数 σ^2 的估计为

$$\hat{\sigma}^2=\frac{S_e}{n-q-1}, \tag{1.1.7}$$

它是 σ^2 的无偏估计,其中 $S_e=\sum(y_i-\hat{y}_i)^2=\hat{e}^T\hat{e},\hat{e}=Y-X\hat{\beta}$.

定理 1.1.1 $\hat{\beta}$ 是 β 的线性无偏估计,其方差协方差矩阵为

$$D(\hat{\beta})=(X^TX)^{-1}\sigma^2. \tag{1.1.8}$$

定理 1.1.2 $Cov(\hat{e},\hat{\beta})=0$.

定理 1.1.3 当 $Y\sim N(X\beta,\sigma^2I)$ 时,$\hat{\beta}$ 与 S_e 相互独立,且 $\hat{\beta}\sim N(\beta,\sigma^2(X^TX)^{-1})$,$S_e/\sigma^2\sim\chi^2(n-q)$,其中 q 为矩阵 X 的秩.

定理 1.1.4 (Gauss-Markov 定理) 在假定 $E(Y)=X\beta,D(Y)=\sigma^2I$ 时,β 的任一线性函数 $c^T\beta$ 的最小方差线性无偏估计(best linear unbiased estimator, BLUE)为 $c^T\hat{\beta}$,其中 c 是任一 q 维列向量,$\hat{\beta}$ 是 β 的最小二乘估计.

该定理的证明可参见文献[3],其他定理的证明可参见文献[7].

非病态线性模型回归系数 β 的最小二乘估计 $\hat{\beta}$ 具有的这些性质,为其研究提供了方便,也正是因为这些性质,才能较容易地获得非病态线性模型的有关显著性检验和异常点的检验函数等,从而也较容易地建立该模型下的有关理论.

1.1.3 非病态线性模型研究的主要内容及其研究方法

文献[1]结合线性模型较为详细地介绍了统计诊断的原理和方法,其主要内容包括异常点检验、影响分析及数据变换等,其中,影响分析包括总体影响分析和局部影响分析,是统计诊断中十分活跃的部分,而总体影响分析是影响分析中方法最直观、理论最实用的内容.

根据文献[1]研究线性模型的基本思想和方法可知,获取模型的有关参数估计,尤其是获取回归系数的一个较优良的估计是研究模型的基础. 回归系数的估计在模型研究中起着其他参数估计不能替代的重要作用,原因在于:第一,在总体影响分析中,通过数据删除模型考察各数据点对统计推断的影响大小,一般都是针对回归系数的估计而言的;第二,衡量既定模型与某一给定的数据集的拟合情况一般都是通过模型本身的参数估计量来体现的,而这些参数估计都与回归系数估计有着直接或间接的关系;第三,回归系数的估计能够唯一地确定回归方程,没有它,作为模型的最终归宿——预测和控制就失去了依据.

非病态线性模型的基本理论是我们研究病态模型的基础,它的研究方法为我们研究病态模型提供了思想方法和模式.

1.2 病态模型概述

病态模型是人们在长期的实践应用中所提炼出来的一种新模型，是对非病态模型的补充和发展．由于在实际中，病态模型要比非病态模型广泛得多，因此研究这种设计阵呈病态的模型更具有深远的意义．

1.2.1 病态模型的产生背景及其存在的意义

任何理论都来源于实践，又可用应用于实践．病态模型(morbid model)是人们对非病态模型认识深化的结果，也是人们的认识更加接近自然的反映．病态模型实际上是对线性模型在其设计阵为非退化阵方面的推广，是对事物联系普遍性的反映．

文献中[6]提到的协方差阵的行列式为零，是对线性模型在满足方差齐性条件方面的推广．文献[3]对多重共线性、自相关性进行了定义，并对其所带来的危害、识别方法及处理措施等进行了讨论，其中的多重共线性指的就是设计阵为退化阵的情形．在定积分中，把积分区间从有限推广到无限，把被积函数从有定义推广到无定义而获得了广义积分．不论是模型的推广，还是定积分的推广，都是为了适应实践发展的需要，都有其存在的实际背景．本书所指的病态模型主要是针对设计阵而言的．

1.2.2 病态模型的一般形式及其特征

1. 病态模型的一般形式

为了定义病态模型，先给出有关的定义

定义 1.2.1 设 X 是一 $n\times q$ 阶阵，如果

$$|X^TX|=0, \tag{1.2.1}$$

则称 X 为退化阵；如果

$$|X^TX|\approx 0, \tag{1.2.2}$$

则称 X 为趋于退化阵．

退化阵与趋于退化阵的区别在于退化阵为列降秩的，且具有明确的秩，而趋于退化阵没有秩可言，它在形式上仍可能是列满秩的．退化阵与趋于退化阵的共同点在于它们都使相应模型回归系数的最小二乘估计分量有偏大的趋势．

退化阵与趋于退化阵分别对应于变量间完全线性相关与近似线性相关．事实上，(1.2.1)式与(1.2.2)式也可叙述如下：

如果 $n\times q$ 阶阵 X 满足(1.2.1)式，则存在不全为零的数 $k_1,k_2,\cdots,k_q$，使

$$k_1X_0+k_2X_1+\cdots+k_qX_{q-1}=0, \tag{1.2.3}$$

或

$$k_1x_{i0}+k_2x_{i1}+\cdots+k_qx_{i,q-1}=0,i=1,2,\cdots,n; \tag{1.2.4}$$

如果 X 满足(1.2.2)式,则有

$$k_1X_0+k_2X_1+\cdots+k_qX_{q-1}\approx 0, \tag{1.2.5}$$

或

$$k_1x_{i0}+k_2x_{i1}+\cdots+k_qx_{i,q-1}\approx 0, i=1,2,\cdots,n. \tag{1.2.6}$$

其中,$X_0=(1,1,\cdots,1)^T$ 表示 X 的第一列,它是分量全为 1 的 n 维列向量,k_i 称为表出系数.

定义 1.2.2 如果 $n\times q$ 阶阵 X 满足(1.2.1)式或(1.2.2)式,则称模型

$$Y=X\beta+\varepsilon,\varepsilon\sim(0,\sigma^2I) \tag{1.2.7}$$

为病态模型.

从实践中提炼出病型模型的目的就是为了对其进行研究,建立有关理论,以便进一步指导实践.

2. 病态模型的特征

病态模型的一大特点是其回归系数的最小二乘估计分量有偏大的趋势. 偏大的本质是不确定,其实质是没有如实地反映客观事实. 如果我们以这样的估计作为基础对病态模型作进一步研究,那么所获得的有关统计推断是不可靠的,由此所建立起来的关于病态模型的理论也是值得怀疑的.

虽然文献[4]通过对病态模型回归系数的分量加以限制的方法获得了其广义岭估计的具体表达式,但这样的岭估计是个向原点压缩的有偏估计. 从衡量一个估计是否优良的标准而言,岭估计不是一个理想的估计. 另外,由于岭估计与模型本身以外的因素约束条件 N 及岭参数 λ 有关,因此变数较大,不易控制. 但是,由于回归系数估计不论是在模型的理论研究,还是在实践中通过回归方程作预测都有重要的作用,因此,要解决有关病态模型的问题,必须首先获得比岭估计更理想的回归系数估计.

1.2.3 病态模型回归系数估计的分量偏大的绝对性与相对性

病态模型回归系数估计的分量有偏大的趋势不是绝对的,而是相对的,也正是因为这一点,病态模型才有其研究的价值. 事实上,病态模型中的"病态"并不是指这样的模型是虚无飘渺的,不可捉摸的,它是实实在在存在的. 病态模型回归系数的最小二乘估计分量有偏大的趋势,也不是因为病态造成的,而是由于我们获取其回归系数估计的方法造成的. 为了说明这一点,我们先来看一个例子.

例 1.2.1 考虑二元回归

$$y_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\varepsilon_i, i=1,2,\cdots,10,$$

其中 $Y=(4,7,10,13,16,19,22,25,28,31)^T$.

相应的设计阵为

$$X=\begin{pmatrix}1 & 1 & 2\\ 1 & 2 & 4\\ 1 & 3 & 6\\ 1 & 4 & 8\\ 1 & 5 & 10\\ 1 & 6 & 12\\ 1 & 7 & 14\\ 1 & 8 & 16\\ 1 & 9 & 18\\ 1 & 10 & 20\end{pmatrix}.$$

显然，设计阵 X 为退化阵，其秩为 2.

由(1.1.6)式可知，此时模型回归系数的最小二乘估计不存在，因为$(X^TX)^{-1}$不存在．所谓不存在，是指回归系数估计的分量为无穷大．而无穷大究竟是多少不明确，不明确实际上表明通过最小二乘法所获得的估计并没有给我们提供一个明确的数量结果，不便于作进一步的统计推断，这样的估计于我们没有任何用处．但是，从我们事先所构造的数据来看，所有数据点都满足关系式：$y=1+x_1+x_2$．这个关系式就应该是该模型的回归(平面)方程．由回归方程的定义知，该模型的回归系数估计当然就是 $\hat{\beta}=(1,1,1)^T$ 了．

综上可知，一方面，我们通过最小二乘法所获得的估计的分量为无穷大；另一方面，我们通过所构造的数据获得其估计为 $\hat{\beta}=(1,1,1)^T$．这说明病态模型回归系数估计的分量趋于无穷不是绝对的，而是相对的．

该例同时也说明了病态模型回归系数的真实估计是存在的．至于通过什么样的方法去获取其真实的估计，正是病态模型值得研究的地方，也是本书所要解决的主要问题之一．

1.2.4 岭估计的有关理论

岭估计是带约束条件线性模型回归系数的最小二乘估计，属于岭估计理论的内容，其表达式为

$$\hat{\beta}(\lambda)=(X^TW^{-1}X+\lambda N)^{-1}X^TW^{-1}Y. \tag{1.2.8}$$

它是考虑到设计阵呈病态时模型回归系数的最小二乘估计的分量有偏大的趋势，从而导致其性质变差，为了改进它的这一性质，通过对其千分量加以约束的方法而获得的估计．岭估计相对于非约束病态模型的最小二乘估计确实有所改进，但并未改变其不唯一性这一性质．

由文献[6]知，岭估计实际上是个向原点压缩的有偏估计，且由于其与约束条件及岭参数有关，因而仍不宜以此为基础对病态模型作进一步研究．通过文献[4]获取其过程可知，约束条件中的 N 及岭参数 λ 都属于模型本身以外的参数．岭估计实际上是

$$S(\beta)=(y-X\beta)^TW^{-1}(y-X\beta) \tag{1.2.9}$$

在约束条件 $\beta^TN\beta=1$ 下的最小值解．根据拉格朗日乘数法知，岭参数是利用拉格朗日乘数法求解目标函数在约束条件下的极值问题时引入的参数．

单从岭估计的表达式(1.2.8)来看,约束条件中的 N 及岭参数 λ 都在影响岭估计 $\hat{\beta}(\lambda)$ 的取值,但事实上它们对岭估计的影响是有差异的,不能等同. 再从(1.2.9)式的结构来看,约束条件中的 N 及岭参数 λ 好像是在一起配合完成使矩阵 $(X^TW^{-1}X+\lambda N)^{-1}$ 有意义这一使命. 如果仅仅从改进病态模型回归系数的最小二乘估计有偏大的趋势这一观点来说,只要 $X^TW^{-1}X+\lambda N$ 可逆就行了.

如果是这样的话,那么岭估计就没有多大的存在价值,只不过是摆设而已. 但事实上,要想从岭估计的表达式(1.2.8)获取我们真正所需要的估计,必须对其加以限制. 当 N 未达到所要求 N 的时,无论岭参数 λ 取什么样的值也得不到我们所需要的估计,而且从中我们还可以知道约束条件中的 N 一般不宜取为纯量阵. 这就是岭估计的奥秘,也是其存在的理由之所在. 但是,在没有通过非约束法获取病态模型的最佳岭估计之前,我们是无法对其存在性给予一个合理的解释的,更无法对其取值进行定量的控制. 由此看来,是非约束法赋予了岭估计新的生命.

我们不能因为通过非约束法获得了病态模型的真正估计,就因此对岭估计加以否定. 事实上,没有岭估计,我们也很难发现获取病态模型回归系数估计的方法. 确切地讲,是岭估计帮助我们解决了病态模型的有关问题. 再者,没有岭估计作为比较,我们又怎么知道通过非约束法所获得的估计就比岭估计更优良呢? 正是因为所获得的岭估计不是很令人满意,因此才促使我们进一步探索病态模型的更优良的估计. 请看下例.

例 1.2.2　续例 1.2.1,有关数据见表 1.2.1.

基于病态模型(1.2.7)式的第 i 个数据点的 Cook 距离定义为:

$$D_i=\frac{[\hat{\beta}(i)(\lambda_i)-\hat{\beta}(\lambda)]^TX^TX[\hat{\beta}_{(i)}(\lambda_i)-\hat{\beta}(\lambda)]}{q\hat{\sigma}^2},$$

其中,岭参数 λ 及诸 λ_i 的确定见后面第七章.

(1)取 $W=I$,$N=I$. 经计算得有关数据,见表 1.2.1 的后三列;

(2)取 $W=I$,$N=diag(1,4,9)$. 经计算得有关数据,见表 1.2.2 的后三列.

当取 $N=I$ 时,由表 1.2.1 的倒数第三列知,各岭参数 λ_i 是不相同的,这说明岭参数 λ_i 随删除的数据点的不同而不同;从其倒数第二列可知,残差都较大;最后一列表明,第 7 号数据点的影响最大,其次是第 10 号点.

当取 $N=diag(1,4,9)$ 时,由表 1.2.2 可知,各岭参数 λ_i 也是不同的;残差明显偏大;最后一列表明第 5 号数据点的影响最大. 另外,所获得的方差估计 $\hat{\sigma}^2=203.3281$ 也明显较大.

以上分析说明:当固定 $W=I$,而取不同的 N 时,所得的影响分析结果完全不同;从所构造的数据来看,所有数据点都满足关系式:$y=1+x_1+x_2$. 这说明所获得的回归方程应该与此相差不大. 但是在两个模型下所得到的回归方程都与 $y=1+x_1+x_2$ 相差较大. 这一点可从所获得的回归系数估计

$$\hat{\beta}(778)=(0.0663,0.4463,0.8926)^T$$

$$\hat{\beta}(748)=(0.1713,0.2956,0.2628)^T$$

看出.