统计诊断理论与陀螺定向数据处理

王同孝　靳奉祥　郑文华　著

中国矿业大学出版社

责任编辑　马跃龙　宋党育
封面设计　肖新生

ISBN 7-81040-783-X

ISBN 7-81040-783-X/P·41 定价:18.00元

统计诊断理论与陀螺定向数据处理

王同孝　靳奉祥　郑文华　著

中国矿业大学出版社

内容提要

本书分为三个部分：统计诊断理论的研究、陀螺经纬仪定向方法及数据处理方法的研究、统计诊断理论用于陀螺经纬仪定向数据处理模型的数据分析和模型分析方法。第一部分介绍了作者几年来在线性模型的统计诊断领域所取得的部分研究成果，如线性模型的几何性质及统计检验，线性模型的扰动影响分析理论等；第二部分介绍了作者几年来在陀螺经纬仪定向技术领域所取得的部分研究成果，如陀螺经纬仪定向方法和数据处理方法等；第三部分是统计诊断理论成果在陀螺经纬仪定向数据处理模型及数据分析中的应用。

统计诊断理论与陀螺定向数据处理

王同孝　靳奉祥　郑文华　著

责任编辑　马跃龙　宋党育

中国矿业大学出版社出版发行

新华书店经销　中国矿业大学印刷厂印刷

开本 850×1168　1/32　印张 7　字数 176 千字

1997 年 12 月第一版　1997 年 12 月第一次印刷

印数 1～2100 册

ISBN　7 - 81040 - 783 - X

P・41　　定价：18.00 元

前　言

统计诊断理论是自70年代以来发展起来的一门统计学分支。它一出现就以其强烈的应用背景、新颖的统计思想、广泛的研究内容和丰富的实际成果展现出一个理论与应用紧密结合的新领域。作为对传统统计理论的补充，该理论已成为一个重要的研究领域，为研究数据的特性和模型的结构，以及两者的关系提供了有效的手段。几年来作者致力于研究和扩展统计诊断理论及其在测量数据处理中的应用，先后对统计诊断中的模型数据扰动和方差扰动理论进行了较为系统的扩展，并将其应用于线性模型的数据与模型分析中，弥补了传统分析方法的不足。值得一提的是作者紧密结合陀螺经纬仪定向数据处理这一应用对象，在对这一对象应用了较多的方法和理论的研究基础上，又对模型本身及数据进行了影响分析，为进一步研究提出了明确的思路和方法。为了能使读者有一个较为全面的了解，本书第二部分提供了有关陀螺经纬仪定向方法及数据处理理论研究的成果。作者借本书将几年来的研究成果奉献给大家，意在抛砖引玉，希望有更多的读者从事该领域的研究。另外，本书中所列出的研究成果是煤炭部科教司(原统配煤矿总公司教育局)发展基金项目“摆幅光学测微系统研制”、山东省青年自然科学基金项目“测绘信息模式处理与辨识理论研究”、煤炭部科教司留学回国人员基金项目“统计诊断理论及其在矿山变形模式中的应用”等项目的成果。另外，独知行讲师和马瑞金助教也参加了本书部分内容的研究与编写工作。

由于我们的水平有限，难免存在错误与不足之处，恳请各位专家及读者惠于指正。

作　者

1997年8月

目　录

第一部分　理论基础

第二部分　陀螺经纬仪定向方法及数据处理

第三部分　定向数据处理模型分析

第一部分

理论基础

1 统计诊断引论

统计诊断是70年代发展起来的一门统计学分支。随着计算机技术的飞速发展，这个新的分支得到了迅速发展与完善，较大多数经典的统计方法，诸如参数估计、假设检验、线性回归、多元分析等显示出较大的优势。经过20多年的发展，统计诊断已形成了较为完善的理论体系，尤其是在应用方面展现出一个理论与应用紧密结合的光辉前景。

统计诊断就是对研究对象实施检测，通过检测的信号（数据）来研究和描述研究对象的特征、规律和性质。通过研究描述模型与客观实体之间、数据与模型之间的内在联系和影响，从而获取正确反映客观实体的数据，最终考查既定模型的有效性、合理性和真实性。作为一种新的手段和方法，它已广泛用于各种工程实践中。本书就统计诊断理论基础和作者在该领域中的某些研究成果做一介绍，并结合陀螺经纬仪定向原理及方法，就定向数据处理的有关问题用统计诊断的原理进行了深入的研究。

1.1 概 述

1.1.1 统计诊断理论及其研究意义

统计学的研究对象是一个数据集 D。为了通过研究数据集 D 而达到研究实际问题的目的，常用的方法是把它纳入某一方便有效的统计模型 M 进行研究。实际上任何统计模型都只能是对客观复杂过程的一种近似描述，总会不可避免地包含某些假定，有时模

型本身就是一种假定。这样就引出了许多问题，如模型 M 能否近似地反映客观实际？它是否与数据集 D 中的数据相一致？D 中是否有个别数据在采集时出现错误？模型本身是否稳定？对这些问题应深入、细致地进行研究以达到揭示客观规律的目的。

统计诊断就是针对上述种种问题而发展起来的一种分析方法。为了克服既定模型与客观实际之间存在的不一致性，通常有两种途径可循，一是稳健统计，二是统计诊断。对于第二种途径有两个方面的意义：其一是判断实际数据是否与既定模型有较大的偏离，并采取相应对策，这是目前统计诊断研究的主要内容；其二是利用实际数据来检核既定模型正确与否，并对模型采取相应的修改或变换。在这个方面的研究更具有实用性，如模式识别理论和系统辨识理论等。一般情况下，如果实际数据中仅有个别点与既定模型偏离较大，这时我们往往肯定模型；相反，如果许多数据点都与既定模型偏差比较大，一般保留既定模型而对数据进行变换。如果达不到理想的状态和效果，则需寻找更为有效的模型，这是一个极为复杂的问题。

现就最常用的线性回归模型说明回归诊断的内容。设有线性回归模型

$$\boldsymbol{L} = \boldsymbol{AX} + \boldsymbol{\Delta} \tag{1.1.1}$$

其中 $\boldsymbol{L}=(l_1,\cdots,l_n)^{\mathrm{T}}$，$\boldsymbol{\Delta}=(\Delta_1,\cdots,\Delta_n)^{\mathrm{T}}$，$\boldsymbol{X}=(x_1,x_2,\cdots,x_t)^{\mathrm{T}}$，$\boldsymbol{A}$ 为 $n\times t$ 阶列满秩矩阵。通常假定其分量 $\Delta_1,\cdots\Delta_n$ 互相独立，数学期望为零，方差具有齐次性，即

$$\boldsymbol{E}(\boldsymbol{\Delta}) = 0 \qquad \mathrm{var}(\boldsymbol{\Delta}) = \sigma^2\boldsymbol{I}$$

其中 σ^2 为未知常数，$\boldsymbol{I}$ 为 n 阶单位矩阵，这时记为

$$\boldsymbol{\Delta} \sim (0,\sigma^2\boldsymbol{I}) \tag{1.1.2}$$

在大多数情况下还假定 $\boldsymbol{\Delta}$ 服从标准正态分布，即

$$\boldsymbol{\Delta} \sim N(0,\sigma^2\boldsymbol{I}) \tag{1.1.3}$$

在传统的统计学研究内容中，通常要解决给定数据集在三个假设下的最优解问题和解的稳定性。如果数据集中的个别数据并不符合既定模型，那么应采用何种方法探测这些数据并进行合理的处理将是统计诊断所研究的主要内容。对于这些点通常称为异常点，识别、判定和检验这些异常点是统计诊断理论主要研究的问题。如果数据集与既定模型之间有很大的或系统性的偏差，应设法修改假设模型，但更常用的方法是保留原模型而对数据进行变换，使之适合假设条件(1.1.1)、(1.1.2)和(1.1.3)式。回归诊断研究的另一项重要内容就是所谓的影响分析。在利用模型求解的过程中，每组数均对统计推断量产生一定的影响，或者说都起一定的作用。但每组数据的影响不尽相同，必须通过统计量定量地刻画数据点影响的大小，从而找出强影响数据点。一般采用的分析方法是利用残差分析和残差图对既定模型的拟合情况进行行之有效的综合分析。

应当注意的是统计诊断中对异常点的处理和模型影响分析没有进行深入细致的研究，而这些问题在实用中恰恰又是非常需要的。例如：在一个数据集中有某些组数据不符合既定模型，即所谓的异常点，用信息论的观点考虑它们所拥有的信息量是很大的，如何充分利用这些数据是非常值得研究的问题。某一组数据对统计推断的影响不仅与数据本身有关，同样也与模型有关。因此，对统计推断影响的因素包含了数据与模型两个面，应对它们进行全面细致的研究。作者在这两个问题上做了大量工作，愿借本书与广大读者交流。

统计诊断的一般工作过程如图 1.1.1 所示。

综上所述，统计诊断主要研究异常点识别、残差分析、影响分析和数据变换等内容。在该领域中，研究最早也最为成熟的应该是线性模型。因此，本书在讨论线性回归诊断的基础上，对原有的某些理论进行了扩展，并将其用于陀螺经纬仪定向数据处理中。

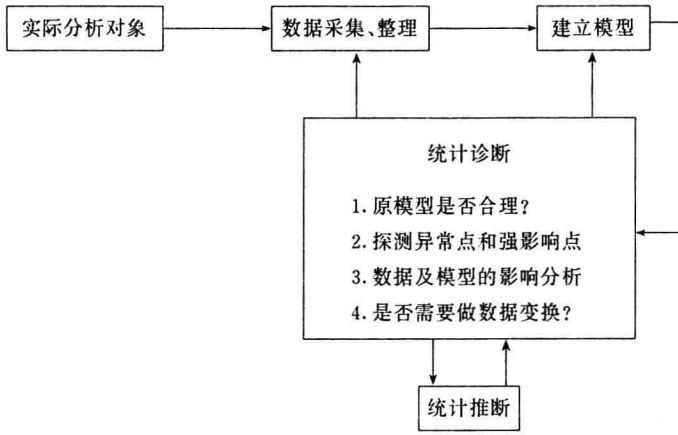

图1.1.1

1.1.2 两个基本概念的解释

统计诊断问题的研究经常涉及两个基本的概念:异常点和强影响点。下面对这两个概念进行必要的解释。

(1)异常点

异常点没有一个确切的定义,通常是指与既定模型偏离很大的数据点。但偏离到何种程度才算大或者才算异常,这就必须对模型误差的分布有一定的了解。目前对异常点有以下两种较为流行的看法:其一,把异常点看成是那些与数据集的主体明显不协调的数据点,这时异常点可解释为所假定的分布中的极端点,即落在分布的单侧或双侧分位点以外的点;其二,把异常点视为杂质点,它与数据集的主体不是来自同一分布,而是来自另一分布的少量杂质。不论哪种看法,所谓异常点的"异常"是相对于数据主体或既定模型而言的。在回归模型中对这种异常点应作较为细致的鉴定,必须通过对度量偏离的指标作检验来确定。

应该注意的是:异常点与数据集或模型所作的分布假定是密

切相关的，不正确的分布假定会导致错误的结论。因此，对异常点的处理必须持慎重态度，不能只是作简单的删除。在许多场合中，异常点的出现恰好是探测某些事先不清楚的或更为重要的分布规律的线索。应当指出的是所有这些工作都必须根据专业知识对数据收集的实际情况进行仔细分析后再作处理。有时会超出统计学研究的范围，此时必须追溯到原始的实际问题。

(2)强影响点

通常数据集中的强影响点是指那些对统计量的取值有非常大的影响力或冲击力的点。在统计推断中，各数据点对统计量的影响大小是不相等的。另外，在分析影响大小时，有几个基本问题需要考虑。首先，必须明确“是对哪一个统计量的影响？”例如，对于线性回归模型，所考虑的是：是对回归系数 X 的估计值 $\hat{X}$ 的影响，还是对误差方差 σ^2 的估计量 $\hat{\sigma}^2$ 的影响；是对拟合优度统计量的影响，还是对预测统计量的影响，等等。分析目标不同，所考虑的影响亦有所不同。一般来讲，对于既定模型，通常总是选择几个有兴趣的统计量(如回归系数估计量等等)，然后再考查数据对它们的影响。其次，必须确定度量影响的尺度是什么。迄今为止已提出多种尺度，诸如基于残差的尺度，基于拟合的尺度，基于影响函数的尺度，基于置信域的尺度和基于似然函数的尺度等。在每一种类型中又可能有不同的统计量，例如基于影响函数就已提出多种“距离”来度量影响，有 Cook 距离、Welsch—Kuh 距离、Welsch 距离、修正的 Cook 距离等等。由此可见，如何研究影响与从何种角度考虑统计问题有密切的关系。每一种度量都着眼于某一方面的影响，并在某种具体场合下较为有效。这一方面反映出度量影响问题的复杂性，另一方面也说明了影响分析的研究在统计诊断领域是一个甚为活跃的方面，还存在大量有待研究解决的问题。作者在本书中将影响分析理论进行了扩展性研究，并将其应用于模式识别与系统辨识中，以对陀螺定向的数据处理模型进行深入而细致的研究。对影响

分析方法的应用可以选择几种不同的度量对影响进行分析，并对各种分析结果加以比较，以期得到更为全面的结论。

对待强影响点也要和对待异常点一样必须慎重处理。强影响点一般是数据集中更为重要的数据点，它往往能提供比一般数据点更多的信息。强影响点和异常点是两个不同的概念，它们之间既有一定的联系也有区别，强影响点可能是异常点，也可能不是；反之，异常点可能是强影响点，也可能不是。

近年来，随着人们对统计诊断理论研究的不断深入以及实际应用的需要，数据点的影响分析已不能满足需要，于是开始研究与数据或模型有关的更一般的因素对于统计分析的影响。因此，影响问题还可以从更一般的观点来考虑——广义影响分析，即研究既定模型有微小扰动时对于统计推断的影响。“扰动”可理解为既定模型所对应的分布有微小变化，其分布从 F 变为 G，而 F 与 G 按分布函数在空间中的某种“距离”非常接近。相应地，统计量 $T(F)$ 看成分布函数空间的泛函，受到了分布 F 扰动的影响而变为 $T(G)$。由此可以研究扰动对于统计量 $T(F)$ 的影响。换言之，需研究 $T(F)$ 对扰动的敏感性或稳健性。在统计诊断中，通常把扰动归结为与模型有关的若干因素所造成的，从而定量地刻划扰动，并提出度量影响的统计量。就线性模型而言，常见的扰动方式有均值的漂移、方差的扩大和自变量的改变等。关于影响的刻划，似然距离受到了很多统计学家的重视，因为它有明确的统计意义，适用范围也十分广泛。近年来，还有人用微分几何观点分析似然距离的变化，利用统计曲率来研究扰动的局部影响。

综上所述，作为统计学的一个迅速兴起的新分支，统计诊断已逐步展现出丰富的研究内容和广阔的应用前景。目前发展比较成熟、应用比较成功的主要还在线性回归诊断方面。在其它统计领域也有人做了大量的研究工作，但多数只是一些初步的探索，在许多方面还留有大片未开垦的“处女地”。人们清楚地知道：统计诊断的

效果和统计分析方法应用的成败是密切相关的。可以预测，在不远的将来，统计诊断会受到更多统计工作者的重视和研究，得到较快的发展。

1.2 线性模型的参数估计

1.2.1 引 言

在测量数据处理中所采用的基本数学模型为 Gauss—Markov（简称 GM）线性模型。该模型在参数估计和统计性质的描述及证明过程中运用了大量的代数公式，这对深刻认识及灵活运用该模型带来很大的不便。为此，本节以空间向量几何理论为基础，采用空间向量方法简明地描绘出了各向量之间的几何及统计关系，实现了 GM 模型的图形化，为灵活运用该模型提供了一种较好的数学方法。关于这个问题陈永奇(1987)进行过研究，但尚未将该模型图形化；崔希璋等(1991)仅用向量投影概念进行了描述；韦博成等(1991)只对 $\boldsymbol{HX}=0$ 约束条件进行了图示分析。前人对该模型的向量化表示与分析缺乏细致性和完整性，其图示亦缺乏严密性，而这正是本文着重研究的内容。

1.2.2 模型的几何分析

设有 GM 线性模型

$$\underset{n\times 1}{\boldsymbol{l}^*} = \underset{n\times 1}{\boldsymbol{u}^*} + \underset{n\times 1}{\boldsymbol{\Delta}^*} = \underset{n\times 1}{\boldsymbol{A}^*}\underset{n\times 1}{\boldsymbol{X}} + \underset{n\times 1}{\boldsymbol{\Delta}^*} \tag{1.2.1}$$

$$\boldsymbol{\Delta}^* \sim N(0,\sigma^2\boldsymbol{Q}_{\Delta}^*) \qquad \boldsymbol{Q}_{\Delta}^* = \boldsymbol{P}^{-1} \tag{1.2.2}$$

式中 $\boldsymbol{l}^*$ 为观测向量；$\boldsymbol{u}^*$ 为 $\boldsymbol{l}^*$ 的期望值向量；$\boldsymbol{A}^*$ 为设计矩阵；$\boldsymbol{X}$ 为未知参数向量；$\boldsymbol{\Delta}^*$，$\boldsymbol{Q}_{\Delta}^*$ 和 $\boldsymbol{P}$ 分别为随机误差向量、误差向量的协因数阵和权阵；σ 为单位权方差。

令 $\boldsymbol{P}=\boldsymbol{g}\boldsymbol{g}^{\mathrm{T}}$($\boldsymbol{g}$ 为三角阵),设 $\boldsymbol{l}=\boldsymbol{g}^{\mathrm{T}}\boldsymbol{l}^{*}$,$\boldsymbol{A}=\boldsymbol{g}^{\mathrm{T}}\boldsymbol{A}^{*}$,$\boldsymbol{u}=\boldsymbol{g}^{\mathrm{T}}\boldsymbol{u}^{*}$,$\boldsymbol{\Delta}=\boldsymbol{g}^{\mathrm{T}}\boldsymbol{\Delta}^{*}$,则模型(1.2.1)、(1.2.2)变为等价模型

$$\boldsymbol{l}=\boldsymbol{u}+\boldsymbol{\Delta}=\boldsymbol{A}\boldsymbol{X}+\boldsymbol{\Delta} \tag{1.2.3}$$

$$\boldsymbol{\Delta}\sim N(0,\sigma^{2}\boldsymbol{I}) \tag{1.2.4}$$

期望值向量 $\boldsymbol{u}=\boldsymbol{A}\boldsymbol{X}$ 的定义域为 $\mathbf{R}^{n}$ 中的 t 维线性子空间 $\Omega=S\{\boldsymbol{A}\}$(由 $\boldsymbol{A}$ 的列向量所张成的空间)。设 $\boldsymbol{J}$ 为该子空间的投影阵,$\boldsymbol{J}_{\perp}$ 为其正交补空间的投影阵,则有

$$\boldsymbol{J}=\boldsymbol{A}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}} \tag{1.2.5}$$

$$\boldsymbol{J}_{\perp}=\boldsymbol{I}-\boldsymbol{J}=\boldsymbol{I}-\boldsymbol{A}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}} \tag{1.2.6}$$

利用空间向量可以将各向量的最小二乘解向量用图 1.2.1 的形式表示出来。

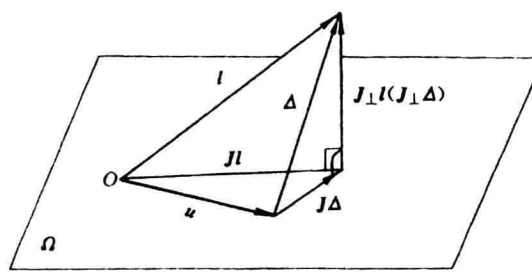

图 1.2.1

参数及观测向量的最小二乘估计分别为

$$\hat{\boldsymbol{X}}=(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{l}=\boldsymbol{N}^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{l} \tag{1.2.7}$$

$$\hat{\boldsymbol{u}}=\boldsymbol{J}\boldsymbol{l}=\boldsymbol{A}\boldsymbol{N}^{-1}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{l} \tag{1.2.8}$$

式中 $\boldsymbol{N}=\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}=(\boldsymbol{A}^{*})^{\mathrm{T}}\boldsymbol{P}\boldsymbol{A}^{*}$。单位权方差估计为

$$\begin{aligned}\hat{\sigma}^{2}&=(n-t)^{-1}RSS=(n-t)^{-1}\hat{\boldsymbol{V}}^{\mathrm{T}}\hat{\boldsymbol{V}}\\&=(n-t)^{-1}\boldsymbol{l}^{\mathrm{T}}\boldsymbol{J}_{\perp}\boldsymbol{l}\end{aligned} \tag{1.2.9}$$

式中 $\hat{\boldsymbol{V}}=-\boldsymbol{J}_{\perp}\boldsymbol{l}=-\boldsymbol{J}_{\perp}\boldsymbol{\Delta}=\boldsymbol{A}\hat{\boldsymbol{X}}-\boldsymbol{l}$(改正数向量)。