测量误差对统计推断的影响

席雷[1], 吴诚鸥[1], 吴令云[2]
(1. 南京信息工程大学 数理学院, 江苏 南京 210044; 2 南京审计学院 信息科学学院, 江苏 南京 211815)

摘要: 讨论了观测误差对某些统计推断的影响, 给出了减小随机误差影响的方法。
关键词: 随机误差; 判别分析; 马氏距离; 自助法
中图分类号: O213.9 文献标识码: A 文章编号: 1000-2022(2007)03-0428-05

Effect of Metrical Error on Statistical Inference

XI Lei[1], WU Cheng-ou[1], WU Ling-yun[2]
(1. School of Mathematics and Physics, NUIST, Nanjing 210044, China;
2 School of Information Science, NAU, Nanjing 211815, China)

Abstract The influence of observational error on statistical inference is discussed and the method to diminish the impact of stochastic error is given.
Key words stochastic error; discriminant analysis; Mahalanobis distance; bootstrap

0 引言

由于测量设备计量性能的局限性、周围环境的影响、模拟式仪器的读数存在人为偏差以及测量方法和测量程序的近似性等原因, 观测值和真值之间总是存在一定的差异, 在数值上即表现为测量误差。传统的减小测量误差的方法主要是改进测量精度、改善测量方法, 并通过获取尽可能多的误差信息和利用计算机软件来补偿等方法降低误差[1-2]。相关研究的方法有: 测头半径补偿 (包括微平面法, 平均向量法, 三点共圆取线法等方法)、测头校正以及减小零件坐标系产生的误差等。

近年来, 具有观测误差的线性与非线性回归模型受到重视, 是当前热门课题[3-7]。在一些线性回归模型中, 有的采用研究的是多次测量结果和偶然误差的估计方法以及一次测量结果误差的估计方法。而非线性模型中采用包括稳健估计、有偏估计及相关抗差有偏估计等方法[8-11]。本文举例说明测量误差会改变大 T 方检验和判别分析的结果, 进而造成错误, 并且研究了测量误差影响假设检验和判别分析的原理, 同时也提出避免测量误差造成错误的方法。

1 实例与数学模型

例 **1** Johnson等[12]举出健康成年女性汗液有关数据, 并用大 T 方检验判断其总体均值是否等于 $(4, 50, 10)^T$, 计算结论是: 总体均值等于 $(4, 50, 10)^T$。表 1给出健康成年女性汗液的模拟数据, 该数据四舍五入即得文献[12]的数据 (具有测量误差的数据)。也就是说, 假定真实数据是表 1中的数据 (其中 X_1、X_2、X_3 为影响因子), 用表 1的数据作大 T 方检验, 推断出总体均值不等于 $(4, 50, 10)^T$, 与文献[12]的结果相反, 可见测量误差改变了推断结果。

例 **2** 文献[13]列出长江中游 1951—1975年 6月降水等级的数据, 并用最大概率判别法 (实质上是马氏距离判别)进行判别, 回代时 3个数据有误差, 用它们判断 1976—1979年的降水等级, 也有 3个误判。现模拟表 2中 1961和 1979年 2 a的气象因子模拟数据, 即 $X_1 \sim X_4$ ($X_1 \sim X_4$ 为判别雨量的 4个因子)分别模拟为 0.423、83.73、38.11、20.44及 0.429、83.55、33.99、60.54。数据四舍五入后可得文献[13]这两年的数据; 即假定文献[13]中数据具有测量误差, 模拟后的数据是真实值。由表 2数据用最大概率判别法判别, 1951—1975年回代只有 2

收稿日期: 2006-05-29; 改回日期: 2006-12-30
作者简介: 席雷 (1981-), 安徽蚌埠人, 硕士, 研究方向: 数理统计中测量误差理论, ljhsnk@sohu.com.

表 1　汗液的模拟数据

Table 1　simulated data of sweat liquor

序号	X_1	X_2	X_3	序号	X_1	X_2	X_3	序号	X_1	X_2	X_3
1	3.749	48.45	9.349	8	7.249	33.05	7.649	15	1.549	13.45	10.149
2	5.749	65.05	8.049	9	6.749	47.35	8.549	16	8.549	56.35	7.149
3	3.849	47.15	10.949	10	5.449	54.05	11.349	17	4.549	71.55	8.249
4	3.249	53.15	12.049	11	3.949	36.85	12.749	18	6.549	52.75	10.949
5	3.149	55.45	9.749	12	4.549	58.75	12.349	19	4.149	44.05	11.249
6	4.649	36.05	7.949	13	3.549	27.75	9.849	20	5.549	40.85	9.449
7	2.449	24.75	14.049	14	4.549	40.15	8.499				

个数据有误差, 判别 1976—1979的降水等级也只有 2个误判。可见这种情况下, 测量误差使判别结果变坏了。

例 1与例 2不是偶然现象, 分析例 2的计算过程, 以此来说明测量误差对实际问题中的判别分析可能存在影响。将长江中游 6月降水分为 3级: 偏少、偏多、正常, 分别用 A、B、C来表示这 3个母体, 以 $X_1 \sim X_4$ 为判别雨量的 4个因子, 表 2为 1951—1979年的观测资料。

表 2　1951—1979年气象因子 $X_1 \sim X_4$ 的观测值

Table 2　Observed values of meteorological factors X_1 to X_4 from 1951 to 1979

年份	X_1	X_2	X_3	X_4	原分类	年份	X_1	X_2	X_3	X_4	原分类
1951	0.58	82	44	40.6	A	1966	0.65	81	31	28.9	A
1952	0.40	83	18	43.0	B	1967	0.66	83	38	46.6	A
1953	0.55	85	36	30.7	B	1968	0.53	80	42	93.1	C
1954	0.40	85	36	40.7	B	1969	0.56	85	18	16.3	C
1955	0.48	88	49	43.0	B	1970	0.45	83	37	23.9	C
1956	0.41	82	35	78.6	C	1971	0.34	80	42	26.3	C
1957	0.65	80	29	33.2	A	1972	0.41	79	38	40.8	C
1958	0.45	82	32	33.1	C	1973	0.53	83	23	61.3	C
1959	0.39	81	27	46.5	C	1974	0.48	84	19	23.0	B
1960	0.34	85	28	41.7	C	1975	0.30	85	27	17.5	B
1961	0.42	84	38	20.4	C	1976	0.42	81	21	52.2	C
1962	0.52	86	38	0.2	A	1977	0.52	81	38	45.8	A
1963	0.46	88	25	56.7	B	1978	0.36	82	34	34.9	B
1964	0.48	83	46	13.6	A	1979	0.43	84	34	60.5	C
1965	0.53	84	41	32.3	A						

表 2中数据包含测量误差, 其绝对值分别小于 0.005、0.5、0.5和 0.05。为了进行判别, 计算 4个因子在 3种降雨级别的协方差阵, 分别得到下面 3个矩阵:

$$\Sigma^{(1)} = \begin{bmatrix} 0.0 & 0.3 & -0.2 & 0.4 \\ 0.3 & 4.2 & -10.1 & -19.5 \\ -0.2 & -10.1 & 61.8 & 38.7 \\ 0.4 & -19.5 & 38.7 & 608.9 \end{bmatrix},$$

$$\Sigma^{(2)} = \begin{bmatrix} 0.01 & 0.07 & 0.18 & 0.26 \\ 0.07 & 4.57 & 9.29 & 12.17 \\ 0.18 & 9.29 & 106.57 & 18.99 \\ 0.26 & 12.17 & 18.99 & 154.80 \end{bmatrix},$$

$$\Sigma^{(3)} = \begin{bmatrix} 0.0 & 0.0 & -0.2 & 0.3 \\ 0.0 & 3.9 & -7.0 & -14.7 \\ -0.2 & -7.0 & 62.4 & 27.5 \\ 0.3 & -14.7 & 27.5 & 532.4 \end{bmatrix}。$$

这 3个矩阵的元素也包含测量误差, 但表中测量误差对 3个矩阵中数的影响有多大? 这些数含有多大的误差? 为了直观简便, 用有效数字的计算来粗略估计测量误差的影响, 且仅估计有关第 2个因子 X_2 的数据: 表面上看 X_2 有 2个有效数字, 而 $\overline{X_2}$ 是由平均而得, 它的精度较高。所以对于 X_2, $X_2^{(k)} - \overline{X_2}$ 精确到个位, 即误差绝对值小于 0.5, 而 $X_2^{(k)} - \overline{X_2}$ 一般只有 1位整数, 因而只有 1个有效数字; 做乘法运算所得的积仍保持 1个有效数字; 从而 3个矩阵的第 2行和第 2列只有 1个有效数字, 用它们作算

术运算, 有效数字可能进一步减少, 计算的可靠性当然有问题。于是带有与不带有测量误差的判别分析之间可能有较大的差别。

由此可见, 用带有测量误差的数据作统计推断, 除了要考虑一般统计推断的错误 (例如第 1类错误、第 2类错误、计算中舍入误差等) 外, 还应当考虑测量误差的影响。下文将对测量误差的影响, 特别是对假设检验和判别分析的影响做初步讨论。

建立测量误差对统计推断影响的数学模型。

定义 **1** 设已有统计分析问题 H 和真数据 (m 维随机向量) 集 $X_1, X_2, \cdots, X_n$, 舍入误差准则设为 RR, $X_1, X_2, \cdots, X_n$ 按准则 RR 变成 $Y_1, Y_2, \cdots, Y_n$; $H(X_1, X_2, \cdots, X_n)$ 与 $H(Y_1, Y_2, \cdots, Y_n)$ 结论的不一致性称为舍入误差准则 RR 的影响。

为了便于计算和讨论, 改用定义 2。

定义 **2** 设已有统计分析问题 H 和舍入后数据 (m 维随机向量) 集 $Y_1, Y_2, \cdots, Y_n$, 观测误差集 (m 维随机向量) $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$, 其中 $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ 是独立同分布的 (根据测量学惯例, 如不另外声明, 一般为在 $[a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_m, b_m]$ 上均匀分布)。$H(Y_1, Y_2, \cdots, Y_n)$ 与 $H(Y_1 + \varepsilon_1, Y_2 + \varepsilon_2, \cdots, Y_n + \varepsilon_n)$ 结论的不一致性称为舍入误差准则 RR 的影响。

2 消除测量误差影响的方法

由例 1和例 2可见, $H(Y_1, Y_2, \cdots, Y_n)$ 与 $H(Y_1 + \varepsilon_1, Y_2 + \varepsilon_2, \cdots, Y_n + \varepsilon_n)$ 结论不一定一致, 这使得统计推断的正确性增加疑问。为了保证统计推断的可靠性, 采用如下方法。

(1)提高观测数据的精确度。从例 2中的分析可见, 尤其应当检查数据减去均值后的有效数字是否够用。但是提高观测精度要耗费可观的资金, 例如人体温度的测量精度由小数点后一位提高到小数点后两位, 要更换现有温度计, 将耗费可观的资金。有时甚至无法提高观测精度: 例如所观测的是人体某处的温度, 每个人全身各处温度是不完全一致的, 因而过分提高测量人体温度的精确度不可行。

(2)分析模型的稳定性。有些统计推断问题 (例如假设检验和判别分析) 从数学上看是函数值分布问题, 某个函数值在不同区域, 就有不同结论。如果测量误差能使函数值改变区域, 就可能影响统计推断结果。因此, 对于一个统计推断问题, 应当选择合适的变量和精度, 使测量误差尽量不影响统计推断的结果。以下就对假设检验和判别分析分别讨论。

定义 **3** 对于假设检验问题 (及相应含测量误差的数据), 若考虑测量误差后, 检验统计量变化范围包含阈值, 则称该假设检验与测量误差有关, 反之, 则称该假设检验与测量误差无关。

假设检验与测量误差是否有关可用如下方法决定: 假设检验的实质是, 计算统计量 $d(x)$, 若 $d(x) \geqslant c$, 则否定 H_0, 接受 H_1。由测量数据的精确度, 可以确定测量误差的取值范围, 从而得到精确数的变化范围, 从而得到 $d(x)$ 的变化范围: 若 c 不在 $d(x)$ 变化范围内, 则该假设检验与测量误差无关, 无需考虑观测误差的影响, 反之则有关, 要考虑测量误差的影响。具体计算时, 由于测量误差很小, 可将 $d(x)$ 近似为线性函数: 设 x_0 是近似数, x 是精确数, 计算 $d(x)$ 的梯度 ∇d。于是 $d(x) \approx d(x_0) + (x - x_0)^T \nabla d$。把每个测量误差取到误差最大限, 一次让其符号与 ∇d 相同, 另一次相反, 就能得到 $d(x)$ 变化的最大幅度。若对于观测数据, c 不在这样算出的变化范围内, 则无需考虑观测误差, 假设检验是稳定的; 否则假设检验是不稳定的。

在例 1中, 可取 $d(x) = (x - \mu_0)^T S (x - \mu_0)$, 其中 $\mu_0 = (4, 50, 10)^T$。当显著水平 $\alpha = 0.10$ 时, $c = 2.44 \times 19 \times 3/(17 \times 20) = 0.409$; 当显著水平 $\alpha = 0.05$ 时, $c = 3.2 \times 19 \times 3/(17 \times 20) = 0.536\ 5$; 而对于包含测量误差的观测数据[12], $d(x) = 0.487$, 当显著水平 $\alpha = 0.10$ 时接受 H_0。而 $d(x) = (x - \mu_0)^T S (x - \mu_0)$ 对 60个观测值的梯度如下 ($x_1 \sim x_3$ 为 3组梯度):

x_1: 0.080 048 3, 0.078 456 4, 0.062 566 1, 0.080 224 4, 0.105 218 6, 0.044 583 9, 0.024 353 7, 0.019 126, 0.014 920 1, 0.036 881 5, 0.024 951 4, 0.059 534 6, 0.037 949 9, 0.051 484, 0.052 125 3, 0.002 985, 0.117 963 8, 0.011 974 1, 0.046 927 6, 0.022 106 7;

x_2: −0.007 141, −0.006 995, −0.005 572, −0.007 157, −0.009 401, −0.003 958, −0.002 14, 0.001 761 7, −0.001 295, −0.003 265, −0.002 194, −0.005 3, −0.003 362, −0.004 577, −0.004 64, −0.000 222, −0.105 44, −0.001 031;

x_3: 0.027 131, 0.026 580 2, 0.021 210 6, 0.035 652 9, 0.015 124 2, 0.008 272 7, −0.006 442, 0.006 442, 0.005 085 7, 0.012 514 1, 0.008 48, 0.020 864, 0.012 878 3, 0.017 459 9,

0.017 675 5, 0.010 501, 0.039 962 9, 0.004 085 7, 0.015 927, 0.007 520 7。

于是给 x_1 每个观测值增加 0.05, 给 x_2 第 8个观测值增加 0.05, 其余每个观测值减少 0.05; x_3 第 7个观测值减少 0.05, 其余每个观测值增加 0.05。这时 $d(x)$达到最大值 0.556 958 3。它大于 0.409 和 0.536 5, 即 $d(x)$的范围包含 0.409和 0.536 5。所以, 例 1是与测量误差有关的。

距离判别的实质是: 计算样品 x 到各总体的某种距离, 将样品 x 判归距离最近的一类。

若对某个样品, 去除测量误差后, 它到某两类距离的远近程度, 与包含测量误差后, 它到某两类距离的远近程度相反, 测量误差就会改变判别的结果, 由此给出定义 4。

定义 4 对于某个距离判别问题 (及相应含测量误差的数据), 若考虑测量误差后, 被判别点到各总体距离远近顺序会改变, 则称该判别问题与测量误差有关; 反之, 则称该假设检验与测量误差无关。

距离判别是否与测量误差有关的计算, 类似于假设检验是否与测量误差有关的计算, 可依据相关结论得出。

3 自助法 (Bootstrap)[10-11]

对一个未知分布总体, 从中抽取 n个样本, 要估计总体某个参数 θ, 用样本估计 θ_n 去代替 θ就会产生误差, 尤其是 n 较小时误差会更大。一般的估计值的期望与实际值相差一个偏差和一个无穷小量。

在误差理论中, 同一量值进行等精度的重复测量得到的测量列 $(X_1, X_2, \cdots, X_n)$, 设 θ_n 为样本容量为 n的估计值, θ_i 为在原样本中切去第 i个样本后的估计值, θ_i 的计算方法完全同 θ_n 一样 (i = 1, 2, …, n)。于是有相应的数学期望

$$E(\theta_n) = \theta + \frac{a}{n} + \frac{\varepsilon}{n^2},$$

$$E(\theta_i) = \theta + \frac{a}{n-1} + \frac{\varepsilon}{(n-1)^2}。$$

其中: a 是常数, $\frac{a}{n}$、$\frac{a}{n-1}$是偏差, ε是无穷小量。

若令 $\theta_i^* = n\theta_n - (n-1)\theta_i$, 则有

$$E(\theta_i^*) = nE(\theta_n) - (n-1)E(\theta_i) = n\left(\theta + \frac{a}{n} + \frac{\varepsilon}{n^2}\right) - (n-1)\left(\theta + \frac{a}{n-1} + \frac{\varepsilon}{(n-1)^2}\right) = \theta + \left(\frac{\varepsilon}{n} - \frac{\varepsilon}{n-1}\right)。$$

即 θ_i^* 的期望值等于总体参数 θ加上一个无穷小量。从而可以认为用 θ_i^* 的均值 $\theta^* = \sum_{i=1}^{n}\frac{\theta_i^*}{n}$对总体参数 θ进行估计, 要比样本参数 θ_n 对 θ的估计更加准确。

下面以一组实测数据为例来验证“自助法”。

对例 2中 1952年降水量因子 X_2 作等精度测量 10次得到数据 (表 3), 求测量结果。

一般情况下的数据处理:

(1)求算术平均值 $X = 83.0$, (2)计算残差 $v_i = X_i - X$ (i = 1, 2, …, n) (表 3); (3)求标准差 $\sigma =$

表 3　两种不同方法的计算数据

Table 3　Computational data of the two methods

i	X_i	V_i	V_i^2	切去 X_i 后的平均值	切去 X_i 后 σ_i	$\sigma_i^* = 9\sigma_9 - 8\sigma_i$
1	83.2	0.2	0.04	83.0	0 247	0.211
2	82.8	- 0.2	0.04	83.0	0 247	0.211
3	82.9	- 0.1	0.01	83.0	0 255	0.147
4	83.2	0.2	0.04	83.0	0 247	0.211
5	83.1	0.1	0.01	83.0	0 255	0.147
6	83.3	0.3	0.09	83.0	0 235	0.307
7	82.7	- 0.3	0.09	83.0	0 235	0.307
8	82.6	- 0.4	0.16	83.0	0 215	0.467
9	82.9	- 0.1	0.01	83.0	0 255	0.147
10	83.2	0.2	0.04	83.0	0 247	0.211
	$X = 83.0$	$\sum_{i=1}^{9} V_i = -0.1$	$\sum_{i=1}^{9} V_i^2 = 0.53$			

$\sqrt{\sum_{i=1}^{n} v_i^2/n-1} = \sqrt{0.53/9} = 0.243$; (4)求平均值标准差 $\sigma_X = \frac{\sigma}{\sqrt{n}} = \frac{0.243}{\sqrt{10}} = 0.077$; (5)平均值的极限误差 $\delta limX = \pm t_\alpha \sigma_X = \pm 2.26 \times 0.077 = \pm 0.174$。其中: 查 t分布表可得出 t_α, $\gamma = n - 1 = 9$, $\alpha = 0.05$。由此, 可以得到最后测量结果 $P = X + \delta limX = 83.0 \pm 0.174$。

用“自助法”计算:

前 3步同上, 得到 $\sigma_9 = 0.243$, 令 σ_i 为切去第 i 个数据后求得的标准差, $\sigma_i^* = 9\sigma_9 - 8\sigma_i$ (表 3)。$\overline{\sigma^*} = \sum_{i=1}^{10} \sigma_i^*/10 = 0.213$, 于是求得平均值标准差 $\sigma_X^* = \frac{\overline{\sigma^*}}{\sqrt{n}} = \frac{0.213}{\sqrt{10}} = 0.067$, 这样其极限误差 $\delta limX = \pm 2.26 \times 0.067 = \pm 0.151$, 得最后测量结果 $P = 83.0 \pm 0.151$。

由此可见, 采用“自助法”后最后测量结果的精确度确实有所提高。这个方法的缺点是计算量比较大, 尤其是分析大量气象数据的时候。但采用计算机后, 这种缺点会得到有效的缓解。所以当 n 比较小, 且最后测量结果需要核准时, “自助法”不失为一种较有效的办法。

4 小结

本文分析了观测误差对假设检验与判别分析的影响, 并且建议采用“自助法”减少测量误差, 达到减小误差影响的要求。

参考文献:

[1] 费业泰. 误差理论与数据处理 [M]. 北京: 机械工业出版社, 2004

[2] 周江文. 系统误差的数学处理 [J]. 测绘工程, 1999, 18(2): 1-4

[3] 费文龙, 朱克云, 吴诚鸥. 关于伴随同化方法的误差分析 [J]. 南京气象学院学报, 2001, 24(2): 237-241.

[4] 吴诚鸥, 沈桐立, 朱玉华. 初始场估计共轭同化方法的精确化 [J]. 南京气象学院学报, 1999, 22(1): 61-68.

[5] 吕纯濂, 陈舜华. 具有对数正态分布几何过程的统计推断 [J]. 南京气象学院学报, 2000, 23(1): 36-41.

[6] 吴诚鸥, 沈桐立, 王顺风. 卫星云图资料估计湿度的适时回归方法及其影响诊断 [J]. 南京气象学院学报, 2001, 24(2): 162-164

[7] 岳朝龙, 黄永兴, 严忠. SAS系统与经济统计分析 [M]. 合肥: 中国科技大学出版社, 2003.

[8] Buonaccorsi J P. Measurement error, linear calibration and inferences for means[J]. Computational Statistics and Data Analysis, 1991(11): 239-257.

[9] Carroll R J, Ruppert D, Stefanski L A. Measurement Error in Nonlinear Models[M]. London: Chapman & Hall/CRC, 1995.

[10] Bickel P J, Ritov Y. Efficient estimation in the errors in variables model[J]. Annals of Statistics, 1987, 22(15): 513-540.

[11] 么枕生, 丁裕国. 气候统计 [M]. 北京: 气象出版社, 1990.

[12] Johnson R A, Wichern D W. Applied Multivariate Analysis[M]. 5ed. New Jersey: Printice Hall, Inc, 2003.

[13] 朱道元, 吴诚鸥, 秦伟良. 多元统计分析与软件 SAS[M]. 南京: 东南大学出版社, 1999.

[14] 包海臣. 减少随机误差标准差的新方法——“刀切法” [J]. 职大学报, 2004(2): 20-21.

[15] 陈华豪. 刀切法在林业数据分析中的应用 [J]. 东北林业大学学报, 1995(3): 48-51.

[16] 林景星, 陈丹英. 计量基础知识 [M]. 北京: 中国计量出版社, 2001.