

文章编号:1007-6492(2001)03-0072-03

关于相关系数作拟合优度判据的研究

张建华

(郑州大学工程力学系,河南 郑州 450002)

摘 要:实验数据处理,常用相关系数的大小来表示线性拟合优劣的判据.通过实验例证分析研究,相关系数好不一定直线拟合结果就好,实验结果线性拟合得好不一定相关系数就很大,说明不宜用相关系数作实验结果拟合优度的判据.只有它为回归系数时,或当有一变量不可能用一个固定的标度来测量时,才有实际意义的.

关键词:相关系数;拟合;优度;回归

中图分类号: O 313 **文献标识码:** A

0 引言

物理实验中,用直线拟合回归方程时,为了定量地表征变量 x 和 y 之间线性相关的密切程度的强弱,常用相关系数 r 来定量描述^[1],并且还用来评价回归方程的拟合优度.相关系数 r 计算公式为

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

当 $r=1$ 时,说明变量之间存在完全的正相关关系;当 $r=0$ 时,说明变量之间不存在相关关系;当 $0 < r < 1$ 时,说明变量之间存在着一定程度的正相关关系;当 $r=-1$ 时,说明变量之间存在着完全负相关关系;当 $-1 < r < 0$ 时,说明变量之间存在着一定程度的负相关关系^[2]. r 的变化范围为 $-1 < r < 1$.如果在实验中相关系数 r 达到 0.999,表明变量之间存在着显著的线性关系,各实验点聚焦在一条直线附近,实验结果就达到要求.如果相关系数 r 很小,表明各数据点偏离直线,说明实验数据很分散,无线性关系,实验结果未达到要求.因此在直线拟合处理数据时,要计算出相关系数,并用此作为测量回归拟合优度的判据.但是,在物理实验中,有些实验的相关系数很好,甚至达到 0.9999,可实验结果却不好;有些实验的实验结果很好,可相关系数却很小.特别是在多元线性回归中,及一些非线性函数关系转化成

线性方程来拟合时更是如此.下面通过一些例证来进行讨论说明.

1 例证

例 1^[3] 用传感器测量空气相对压力系数,测得数据如表 1 所示.

表 1 压强与温度的测量数据

Table 1 Data of pressure and temperature

U_p	V_t	U_p	V_t
-1.65	14.83	6.61	17.84
0.01	15.34	8.05	17.84
1.63	15.84	9.54	18.35
3.25	16.32	11.12	18.85
4.85	16.84	13.11	19.47

计算机处理直线拟合结果: $p = a + bt = 0.95961 + 0.02935t$, 相关系数 $r = 0.99985$, 空气相对压力系数 $\alpha_p = 0.0030586$.

由数据处理结果可以看出,相关系数 r 达到了 3 个 9 以上,直线拟合得相当可观,各数据点都在直线上.但求出的空气相对压力系数却偏小,没有达到目的要求.这表明相关系数很好,不一定实验结果就达到了良好.

例 2 在半导体热敏电阻的特性研究中,测得 $R-T$ 数据如表 2 所示.

对 $R = A \cdot e^{(B/T)}$ 两边同取对数: $\ln R \approx \ln A + B/T$, 整理得 $1/T = \ln R/B - \ln A/B$, 对该式进行处理得: $1/T = 2.78 \times 10^{-4} \cdot \ln R + 1.012 \times 10^{-3}$,

收稿日期:2001-05-07;修订日期:2001-06-30

基金项目:河南省自然科学基金资助项目(994062300)

作者简介:张建华(1946-),男,河南省汝南县人,郑州大学副教授,主要从事物理实验方面的教学和研究.

相关系数 $r = 0.99997$, 达到了 4 个 9. 进一步考察拟合优度, 对 $T_i = B/\ln(R_i/A)$ 求均方根偏差得 $S_T = \sqrt{\sum(T_i - T)^2/(n-2)} = 1.720 \text{ K}$, 进而 $2 \cdot S_T = 3.440 \text{ K}$, 此值大于温度计基本误差极限, 说明实验结果并未达到良好.

表 2 温度和电阻的测量数据

Table 2 Data of temperature and resistance

$t/^\circ\text{C}$	R_t/Ω	$t/^\circ\text{C}$	R_t/Ω	$t/^\circ\text{C}$	R_t/Ω
22.0	5165	45.1	2147	71.3	902
25.1	4573	50.2	1778	74.0	835
30.2	3750	54.0	1551	79.5	712
35.0	3114	59.8	1296	84.0	621
40.1	2585	64.0	1130	90.1	529

例 3 动力学法测量弹性模量的实验中, 测量数据如表 3 所示.

表 3 线到端面距 e 与频率 f 的测量数据Table 3 Data of line distance e from end and frequency f

距离 e	频率 f	距离 e	频率 f
0.7	415.02	3.5	406.74
1.2	410.98	4.0	406.70
2.0	408.84	5.0	406.91
3.0	407.47		

用 EXCEL 外推法进行数据处理, 圆杆振动频率的方程拟合用回归方程的变量方次 $f = f_0 + C_1 X + C_2 X^2 + C_3 X^3 + C_4 X^5 + C_5 X^7 + C_6 X^9$, f_0 的最佳值为 405.94, 标准差 $\sigma = 0.012$, 不确定度 $\Delta f = 0.05$, 最大残差为 0.26 Hz. 这些数据都未显著大于频率测量的误差极限, 但相关系数 r 仅为 0.9982, 看似结果不好.

例 4 用恒定电流场模拟静电场的实验中, 测得一等位面数据如表 4 所示.

表 4 坐标点测量数据

Table 4 Data of coordinate

x_i	y_i	x_i	y_i
1.0098	0.0010	-0.9864	0.0410
0.6216	0.7834	-0.5097	-0.8715
0.2285	0.9794	-0.0280	-0.9852
0.2647	0.9535	0.7384	-0.6880

等权近似拟合圆半径, 由于圆方程非线性, 可将其化成线型方程来拟合. 对于已知圆心的近似值 (u, v) , 令 $r = \sqrt{(x_i - u)^2 + (y_i - v)^2}$; $\cos\theta_i = (x_i - u)/r_i$; $\sin\theta_i = (y_i - v)/r_i$, 可得关于半径为 r_0 , 圆心坐标 x_0, y_0 的线性方程为 $r_i = r_0 + (x_0 - u)\cos\theta_i + (y_0 - v)\sin\theta_i$. 取近似圆心 $u = 0.0010, v$

$= 0.0010$, 得 $r_0 = 1.0000$, 半径标准差为 0.0029, 结果较好, 但拟合的相关系数 r 仅为 0.022, 远远小于 1.

2 结论

从以上实验的实例讨论可说明, 即使相关系数达到 0.999 以上, 也不能够就说明测量结果就达到良好. 反过来, 相关系数不好, 也不能说明测量结果未达到要求. 由此可见, 用相关系数的好坏作判据测量结果的好坏是不可靠的. 所以, 不能根据相关系数的好坏轻易作结论. 但相关系数能间接的表示线性拟合的精密度^[1]. 相关系数仅反映 x 和 y 之间的线性相关情况, 它不适用于非线性相关关系. 应该强调, 线性相关系数 $r = 0$ 并不等于 x 和 y 之间不存在相关关系, x 和 y 之间可能存在着曲线相关, 并且可将非线性转换成线性模型处理. 若 x 和 y 之间存在着相关关系, 则在 y 的总变差中有一部分变差可以用 x 的变差来解释, 但是, 若把 r 的数值视为在 y 的总变差中可以用 x 的变差所解释的部分所占的比例, 则有可能形成对真实情况的高估现象, 而测定系数 r^2 才是说明在 y 的总变差中可以用 x 的变差所解释的部分所占的比例的确切指标.

有 $R_{xx} = \sum(y_i - y)^2 \approx (1 - r^2) \sum(y_i - y)^2$ 可知 r^2 的值越大, R_{xx} 就越小, 回归直线与观测数据拟合的也就越好^[4]. 但是, 在一般的情况下, 样本相关系数往往大于总体相关系数. 设总体相关系数的无偏估计量 $\rho = \pm \sqrt{r^2(n-1)-1/(n-2)}$. 换言之, 样本相关系数 r 往往形成对 ρ 的高估, 尤其在样本容量 n 较小时, 更是如此. 所以, 在点估计的情况下, 用 ρ 去估计不失为了解总体的真实情况的较好方法^[2].

塞伯认为^[4], 虽然 $1 - r^2$ 是关于拟合情形的一个有用的量度, 但用相关系数 r 本身作推断却是一种值得怀疑的用法. 他还引用了 Tukey 于 1954 年提出的一种有争议的但又无不道理的说法, 即相关系数在两种情形并且只是在两种情形下才是有实际意义的, 一种情形是它们为回归系数时, 另一种情形是当有一变量不可能用一个固定的标度来测量时^[1]. Warren 于 1971 年对相关系数与回归的关系问题进行了有益的讨论^[4].

参考文献:

- [1] 朱鹤年. 物理实验研究[M]. 北京: 清华大学出版社, 1994. 95-97.

- [2] 周复恭.应用线性回归分析[M].中国人民大学出版社,1989.37-51. 1992.41-47.
- [4] 塞伯 G A F.线性回归分析[M].方开泰,张永光,译.北京:科学出版社,1981.117.
- [3] 丁慎训.物理实验教程[M].北京:清华大学出版社,

Study on Coefficient of Correlation as Goodness - of - fit Criterion

ZHANG Jian - hua

(Department of Engineering Mechanics, Zhengzhou University, Zhengzhou 450002, China)

Abstract: In the experimental data disposal, correlation is often used as Criterion of linear fitting goodness. However, examples from some concrete experiment show that good linear fitting experiment data may not have larger correlation coefficient that is to say, correlation coefficient is not suitable to be a criterion judging goodness of fitting. Only if it is also a regression one coefficient, or there is one variable that can't be measured with a fixed scale, the correlation coefficient is of significant.

Key words: coefficient of correlation; fitting; goodness; regression

(上接 58 页)

- [3] TERZOPOULOS D, QIN H. Dynamic NURBS with geometric constraints for interactive sculpting[J]. ASM Transaction on Graphics, 1994, 13(2): 103 - 135. deformable models[J]. Computer Graph, 1987, 21(4): 205 - 214.
- [5] TERZOPOULOS D, FLEISCHER K. Deformable models [J]. The Visual Computer, 1988(4): 306 - 331.
- [4] TERZOPOULOS D, PLATT J, BARR A H, et al. Elastically

Synchronism Research on Free - formed Surface and Tool Path Generation Technology

YUN Min, YU Yuan, WANG Xiao - chun

(Institute of NC Technology, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: Based on deformable NURBS(D - NURBS) free - formed curve, a kind of space displacement method in kinematics is used to construct the CAD model of tool sweeping body in NC machining. Then the primary geometry curve and surface are provided with the characteristics of physical surface. In this way, it is allowed that the designer can exert many tools based on force with the natural and dividable method to sculpt interactively and design the surface directly. Therefore, the surface can satisfy the kinematics and mechanics limit in the special NC machine easily. At the same time, the new method to design the NC machining tool path of free - formed surface. The technology of kinematics, CAGD and CAD/CAM are integrated in this method, and provide the base for synchronism shape design and manufacture.

Key words: CAD/CAM; space displacement; D - NURBS; tool path