多元线性回归分析的实例研究

张宇山
(广东商学院数学与计算科学系 广东 广州 510320)

[摘 要]通过运用 SAS 统计软件,针对一实际例子,编程建立线性回归模型,并通过方差分析和共线性判断等对模型参数进行检验,调整模型形式,最后得到与原数据比较拟和的模型。

[关键词]SAS;多元线性回归;多重共线性;方差膨胀因子

【Abstract】By making programmings in SAS, this paper sets up three linear regression models based on real-world data. After analyzing their respective variance and estimating multicollinearity of variables, the models are adjusted to be more adaptive to the data.

【Key words】SAS; Linear analysis; Multicollinearity; Variance inflation factor

1.统计软件 SAS 简介

SAS 是美国 SAS 软件研究所研制的一套大型集成应用软件系统,具有完备的数据存取、数据管理、数据分析和数据展现功能。尤其是创业产品—统计分析系统部分,由于其具有强大的数据分析能力,在数据处理和统计分析领域,被誉为国际上的标准软件和最权威的优秀统计软件包,广泛应用于政府行政管理、科研、教育、生产和金融等不同领域,发挥着重要的作用。SAS 系统操作以编程为主,人机对话界面不太友好,系统地学习和掌握 SAS,需要花费一定的时间和精力。但无论从速度或功能等各个方面,SAS 作为专业统计软件中的巨无霸, 现在还很难有什么统计软件足以与之抗衡。

本文使用的是 SAS For Windows 6.12。

2.实例背景

本文使用的数据来源于 www.statitistics.com 中 Industry 的数据资料库。

某成品的密度可作为衡量该产品的指标,它由生产过程中的 5 个变量所决定:

X_1 = 该成品所含水量

X_2 = 该成品生产过程中所包含的重复使用材料数

X_3 = 生产过程的平均温度

X_4 = 烘干炉中的温度

X_5 = 原材料的质量指标

现有 48 组数据如表 1。

3.回归分析

首先对 48 组数据进行简单的描述统计量的计算,见表 2。

3.1 模型(1)

首先从最简单的线性回归模型入手,假设回归模型形式如下:

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\beta_5x_5+\varepsilon \quad (1)$$

表 1

X1	X2	X3	X4	X5	Y	X1	X2	X3	X4	X5	Y
0	800	135	578	13.195	104	75	800	135	550	12.745	103
0	800	135	578	13.195	102	75	800	135	550	12.745	111
0	800	135	578	13.195	100	75	800	135	550	12.745	111
0	800	135	578	13.195	96	75	800	135	550	12.745	107
0	800	135	578	13.195	93	75	800	135	550	12.745	112
0	800	135	578	13.195	103	75	800	135	550	12.745	106
0	800	150	585	13.180	118	75	800	150	595	13.885	111
0	800	150	585	13.180	113	75	800	150	595	13.885	107
0	800	150	585	13.180	107	75	800	150	595	13.885	104
0	800	150	585	13.180	114	75	800	150	595	13.885	103
0	800	150	585	13.180	110	75	800	150	595	13.885	104
0	800	150	585	13.180	114	75	800	150	595	13.885	103
0	1000	135	590	13.440	97	75	1000	135	530	11.705	116
0	1000	135	590	13.440	87	75	1000	135	530	11.705	108
0	1000	135	590	13.440	92	75	1000	135	530	11.705	104
0	1000	135	590	13.440	85	75	1000	135	530	11.705	116
0	1000	135	590	13.440	94	75	1000	135	530	11.705	112
0	1000	135	590	13.440	102	75	1000	135	530	11.705	111
0	1000	150	590	13.600	104	75	1000	150	590	13.835	110
0	1000	150	590	13.600	102	75	1000	150	590	13.835	115
0	1000	150	590	13.600	101	75	1000	150	590	13.835	114
0	1000	150	590	13.600	104	75	1000	150	590	13.835	114
0	1000	150	590	13.600	98	75	1000	150	590	13.835	114
0	1000	150	590	13.600	101	75	1000	150	590	13.835	114

其中 $x_1, \cdots, x_5$ 表示数据中的自变量，y 为表示产品密度的因变量。进行方差分析和参数检验，得到结果见表 3。

从图 2 结果可以看到，决定系数 R-Square = 0.6375，表明方程模拟得并不理想，回归方程的显著性检验 p 值虽然较理想(<.0001)，但回归系数的显著性检验表明除了常数项和 x3 的系数高度显著外，其余系数都不十分显著，特别是 x4 和 x5。另外我们通过绘制残差图，可以看到：

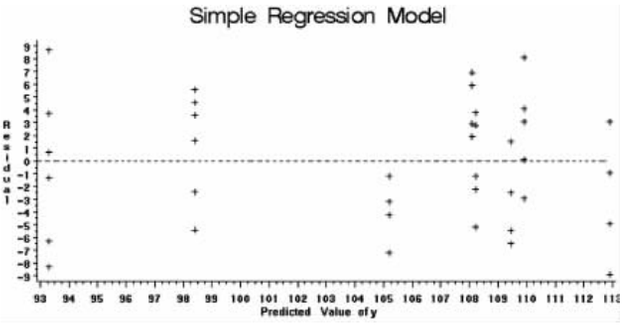

图 1 模型(1)的残差图

可以看到，在 $\hat{y}=105$，残差都在零线以下。所以综合上述，模型(1)并不是一个与原数据拟和十分理想的模型。

表 2

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
X1	48	37.5000	37.8968	1800.0	0	75.0000
X2	48	900.0	101.1	43200.0	800.0	1000.0
X3	48	142.5	7.5794	6840.0	135.0	150.0
X4	48	576.0	22.1004	27648.0	530.0	595.0
X5	48	13.1981	0.6715	633.5	11.7050	13.8850
Y	48	105.7	7.7875	5073.0	85.0000	118.0

表 3

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F-value	Pr>F
Model	5	1817.10716	363.42.143	14.77	< .0001
Error	42	1033.20534	24.60013		
Corrected Total	47	2850.31250			

Root MSE	4.95985	R-Square	0.6375
Dependent Mean	105.68750	Adj R-sq	0.5944
Coeff Var	4.69294		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr>|t|
Intercept	1	134.25774	39.66120	3.39	0.0016
X1	1	0.05032	0.03638	1.38	0.1739
X2	1	-0.01168	0.00726	-1.61	0.1152
X3	1	0.83426	0.14259	5.85	< .0001
X4	1	-0.15384	0.20740	-0.74	0.4624
X5	1	-3.80429	5.87173	-0.65	0.5206

3.2 调整模型的思路

在多元线性回归的应用中可能碰到这样的问题：(1) 在某个检验水平 α 下，整个回归方程的统计检验小于 α，而方程的各偏回归系数的检验却大于 α；(2)根据专业知识，某自变量与因变量间关系密切，但检验结果大于 α。在统计学中这种现象称为多重共线性(multicollinearity)。当自变量间存在近似的线性关系，即某个自变量能近似的由其它自变量的线性函数来描述，就会造成参数估计的误差急剧增大，从而导致了上述的问题。

如何识别自变量组(X1,X2,…,Xn)是否存在多重共线性，首先想到以 X1,X2,…,Xn 中的一个(如 Xi)为因变量，其他的 Xj($j\neq i$)为自变量建立回归方程，看此回归方程的决定系数(记为 Ri^2)是否较大，若 Ri^2 较大，说明 Xi 的变异基本由其它 Xj 的线性回归所决定。对于 n 个自变量相应的就可求得 n 个 Ri^2，只要其中有一个 Ri^2 较大，该组自变量就存在多重共线现象。而在实际中并不需要建立 n 个回归方程来求 Ri^2，而是把方差膨胀因子(variance inflation factor,记为 VIF)作为衡量标准。设 X1,X2,…,Xn 的相关矩阵 CORRx，可以证明 Ri^2 与 $(CORRx)^{-1}$ 的对角线元素 fii 有对应关系：$f_{ii}=(1-R_i^2)^{-1}$

fii 就称为 Xi 的方差膨胀因子，它与 Ri^2 有如下关系：

当 $Ri^2=0$，即 Xi 与其他自变量不线性相关时，VIFi=1；当 $0<Ri^2<1$ 时，VIFi>1；当 $Ri^2=1$，即 Xi 与其他自变量完全线性相关时，VIFi=∞。所有自变量中最大的 VIFi 通常用来作为多重共线性严重程度的指标，如果 max{VIFi}>10，说明共线性可能严重影响了最小二乘估计值，就要进行自变量的筛选等来调整原方程。

在模型(1)中出现了类似于问题(1)的结果，以下为其 VIF 结果。

Variance Variable	DF	Inflation
Intercept	1	0
X1	1	3.63163
X2	1	1.02953
X3	1	2.23149
X4	1	40.14088
X5	1	29.70489

很明显，VIF4 和 VIF5 都太大了。

以下的 Pearson 系数相关矩阵 (Pearson correlation coefficients matrix)反映了各变量之间的关系。

Pearson Correlation Coefficients

Prob > |R| under Ho: Rho=0 / N = 48

	X1	X2	X3	X4	X5	Y
X1	1.00000	0.00000	0.00000	-0.44584	-0.23420	0.51638
	0.0	1.0000	1.0000	0.0015	0.1091	0.0002
X2	0.00000	1.00000	0.00000	-0.04573	-0.07995	-0.10544
	1.0000	0.0	1.0000	0.7576	0.5891	0.4757
X3	0.00000	0.00000	1.00000	0.64018	0.64240	0.32172
	1.0000	1.0000	0.0	< .0001	< .0001	0.0258
X4	-0.44584	-0.04573	0.64018	1.00000	0.95716	-0.33305
	0.0015	0.7576	< .0001	0.0	< .0001	0.0207
X5	-0.23420	-0.07995	0.64240	0.95716	1.00000	-0.26957
	0.1091	0.5891	< .0001	< .0001	0.0	0.0639
Y	0.51638	-0.10544	0.32172	-0.33305	-0.26957	1.00000
	0.0002	0.4757	0.0258	0.0207	0.0639	0.0

由上面的数据可看出，X4 和 X5 有很强的相关性，且 X4 与 y 更相关。由此考虑去除 x4 或把 x4 和 x5 都去除。

3.3 模型(2)和模型(3)

基于上述分析，从 R^2 和 C_p 两方面考虑变量的选择。

统计量 C_p

$$=\left(\frac{\text{具有 p 个参数(包括截距)的子集模型的残差平方和}}{\text{完全模型的误差方差的估计}}\right)-(n-2p)$$

即若用 SSEP 表示 k 个自变量中的 p 个自变量建立的方程的剩余平方和，则

$$C_p=\frac{SSE_p}{MSE}-(n-2p-2)$$

如果每个数对$(p+1, C_p)$表示一个预测变量的子集，则数对(p, C_p)的曲线图显示了预测观察响应的模型的好坏，一般的好的模型其$(p+1, C_p)$点靠近 45 度直线。也就是按照 C_p 准则选择除完全模型外 C_p 值与(p+1)最接近的模型。

同时兼顾 R^2 和 VIF 两方面的考虑，编写 SAS 程序反复迭代，得到模型(2)和模型(3)：

模型(2)：$y=\beta_0+\beta_3x_3+\beta_{12}x_1x_2+\beta_{23}x_2x_3+\beta_4x_4^2+\varepsilon$ (2)

该模型是在所有剔除 X5 后由 $X_1, X_2, X_3, X_4, X_1^2, X_2^2, X_3^2, X_4^2$，以及 X_1X_2，…，X_3X_4 组成的所有可能的自变量组所建立的回归方程中选择出来的。

模型(3)：$y=\beta_0+\beta_2x_2+\beta_{23}x_2x_3+\beta_{13}x_1x_3+\beta_{123}x_1x_2x_3+\varepsilon$ (3)

该模型是从将 x4 和 x5 都剔除了的回归方程中选取出来的。

但是模型(3)的 VIF 还是明显偏大，如下：

Variable	DF	Variance Inflation
INTERCEP	1	0.00000000
X2	1	6.38480145
X13	1	82.00000000
X23	1	5.41736289
X123	1	82.99728629

而模型(2)则符合要求：

Variable	DF	Standardized Estimate	Variance Inflation

INTERCEP	1	0.00000000	0.00000000
X3	1	0.88567868	2.33109818
X1X2	1	0.21251572	1.60950648
X2X3	1	−0.17327640	1.23935515
X4*X4	1	−0.76078293	2.72263588

所以模型(2)比模型(3)好,以下为模型(1)和(2)的 R^2 和 MSE 的比较。

MODEL	R-Square	Root MSE
Model(1)	0.6375	4.95985
Model(2)	0.6964	4.81775

可看出,模型(2)优于模型(1),由模型(2)的残差图也说明了这一点。

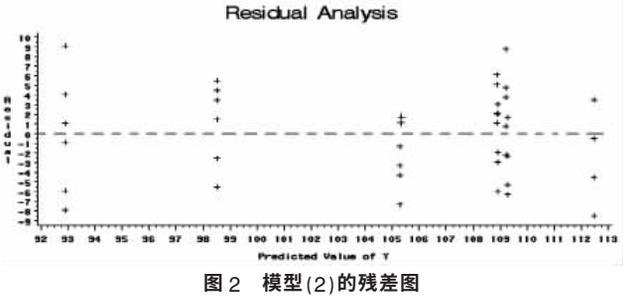

图 2 模型(2)的残差图

至于模型的回归诊断,以下为残差的 Q-Q图。图中的"+"号表示标准正态 u 值的参考直线,"*"号表示实际残差数据点,如果残差服从正态分布,则观测值数据"*"构成的直线与参考直线基本重合。在图中,残差值与参考直线基本重合所以可以认为误差服从正态分布。

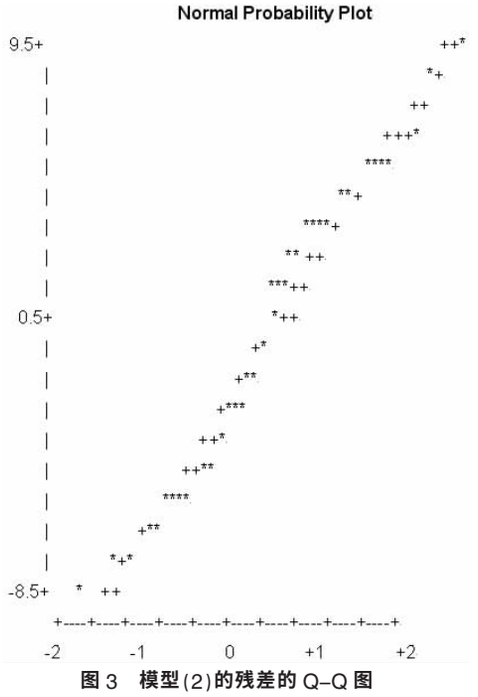

图 3 模型(2)的残差的 Q-Q 图

4.结论

通过三个模型的比较,可以认为模型(2)最好。

把模型(2)的各系数代入方程得:

$$y=64.33222+0.909998x_3+0.000047934x_1x_2-0.000084587x_2x_3-0.000238x_4^2$$

即为所求得的回归方程。

【参考文献】

[1] 王孝仁, 王松桂编译. 实用多元统计分析 [M]. 上海科学技术出版社, 1990,195-264.

[2]上海师范大学数学系概率统计教研组编.回归分析及其试验设计[M].上海教育出版社,1978.

[3]沈其君主编.SAS 统计分析[M].东南大学出版社.

[4]SAS DOC INSIGHT. http://rss.acs.unt.edu/sasdoc/insight.

[5]SAS 6.12 教程. http://medguider.51.net/data/sas/index.htm.

作者简介:张宇山(1975.5—),男,汉族,广东兴宁人,广东商学院数学与计算科学学院讲师,硕士,主要从事应用数学和数理统计的研究。

[责任编辑:韩铭]

(上接第 97 页)找到了这种幻觉。

大学生在找工作时要特别注意:1) 通过正规的招聘单位寻找工作。各高校与教育行政部门所安排的就业招聘需要严格审查单位资质,可靠性比较高。2)通过其他途径找到的工作要严格审查公司的资质与信用,包括从网上、营业地的工商部门查询,要求对方出示营业执照和组织机构代码证书、开户许可证书、税务登记证书和代理授权书。3)"朋友不言商",传销多通过同学、朋友等熟人进行,不要因朋友感情害了自己。4)仔细弄清直销与传销的区别:有无入门费、有无依托优质产品、产品是否在市场上销售、有无退货保障制度、销售人员结构有无超越性、有无店铺经营。

随着社会、经济的发展,新形势下的大学生安全问题也出现了一些新的特征。我们一方面要从现实之中多加观察,提高警惕性、警觉性,对于大学生的安全教育工作严抓不懈,做到"防患于未燃";另一个方面又要能很好的处理突发事故,一旦发生,学校与学生都能沉着应对,采取相应的策略。

[责任编辑:田瑞鑫]