第9章　统计模型

9.1　软件开发人员的薪金

数学建模的基本方法

机理分析　　测试分析

由于客观事物内部规律的复杂及人们认识程度的限制，无法分析实际对象内在的因果关系，建立合乎机理规律的数学模型.

通过对数据的统计分析，找出与数据拟合最好的模型.

回归模型是用统计分析方法建立的最常用的一类模型.

- 不涉及回归分析的数学原理和方法 .
- 通过实例讨论如何选择不同类型的模型 .
- 对软件得到的结果进行分析，对模型进行改进.

9.1 软件开发人员的薪金

建立模型研究薪金与资历、管理责任、教育程度的关系.

分析人事策略的合理性，作为新聘用人员薪金的参考.

46名软件开发人员的档案资料

编号	薪金	资历	管理	教育	编号	薪金	资历	管理	教育
01	13876	1	1	1	42	27837	16	1	2
02	11608	1	0	3	43	18838	16	0	2
03	18701	1	1	3	44	17483	16	0	1
04	11283	1	0	2	45	19207	17	0	2
…	…	…	…	…	46	19346	20	0	1

资历~ 从事专业工作的年数；管理~ 1=管理人员, 0=非管理人员；
教育~ 1=中学，2=大学，3=更高程度.

分析与假设

y~ 薪金，x_1~资历（年）

$x_2 = 1$~ 管理人员，$x_2 = 0$~ 非管理人员

教育

1=中学

2=大学

3=更高

$$x_3 = \begin{cases} 1, & \text{中学} \\ 0, & \text{其他} \end{cases} \qquad x_4 = \begin{cases} 1, & \text{大学} \\ 0, & \text{其他} \end{cases}$$

⇨

中学：$x_3=1, x_4=0$；

大学：$x_3=0, x_4=1$；

更高：$x_3=0, x_4=0$

假设资历每加一年薪金的增长是常数；

且管理、教育、资历之间无交互作用.

线性回归模型 $y = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + \varepsilon$

$a_0, a_1, \ldots, a_4$是待估计的回归系数，ε是随机误差

模型求解

$$y = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 + a_4 x_4 + \varepsilon$$

参数	参数估计值	置信区间
a_0	11033	[10258 11807]
a_1	546	[484 608]
a_2	6883	[6248 7517]
a_3	-2994	[-3826 -2162]
a_4	148	[-636 931]
R^2=0.9567	F=226	p<0.0001 s^2=10^6

资历增加1年
薪金增长546

管理人员薪金
多6883

中学程度薪金比
更高的少2994

大学程度薪金比
更高的多148

$R^2, F, p \rightarrow$ 模型整体上可用

x_1~资历(年)

$x_2 = 1$~ 管理,

$x_2 = 0$~ 非管理

中学: $x_3=1, x_4=0$;

大学: $x_3=0, x_4=1$;

更高: $x_3=0, x_4=0$.

a_4置信区间包含零点，解释不可靠!

[b, bint,r,rint,stats]=regress(Y,X,alpha)

回归系数的区间估计

残差

置信区间

用于检验回归模型的统计量，有三个数值：相关系数r^2、F值、与F对应的概率p

显著性水平（缺省时为0.05）

相关系数 r^2越接近 1，说明回归方程越显著；

$F > F_{1-\alpha}(k, n-k-1)$ 时拒绝 H_0，F 越大，说明回归方程越显著；

与 F 对应的概率 $p<\alpha$ 时拒绝 H_0，回归模型成立.

b

5x1 double

	1
	1.1033e+04
	546.1276
	6.8825e+03
	-2.9942e+03
	147.7380

bi

5x2 double

	1	2
	1.0258e+04	1.1807e+04
	484.4486	607.8067
	6.2481e+03	7.5170e+03
	-3.8263e+03	-2.1620e+03
	-635.7184	931.1944

r

46x1 double

	1
1	-1.5912e+03
2	29.1380
3	239.6051
4	-443.5999
5	188.1380
6	1.7167e+03
7	-500.7276
8	1.4042e+03
9	70.0104
10	-505.8552
11	-1.5845e+03
12	1.6696e+03
13	246.3498
14	1.1939e+03
15	163.2222
16	13.7551
17	-480.9829

ri

46x2 double

	1	2
	-3.4567e+03	274.2766
	-1.9329e+03	1.9911e+03
	-1.7448e+03	2.2240e+03
	-2.4199e+03	1.5327e+03
	-1.7730e+03	2.1492e+03
	-172.1302	3.6056e+03
	-2.4891e+03	1.4877e+03
	-521.7364	3.3301e+03
	-1.8999e+03	2.0399e+03
	-2.5056e+03	1.4939e+03
	-3.4762e+03	307.2432
	-235.6824	3.5749e+03
	-1.7519e+03	2.2446e+03
	-766.9598	3.1548e+03
	-1.8399e+03	2.1663e+03
	-1.9662e+03	1.9937e+03
	-2.4908e+03	1.5288e+03

s

1x4 double

	1	2	3	4
1	0.9567	226.4258	2.3110e-27	1.0571e+06
2				

结果分析　残差分析方法

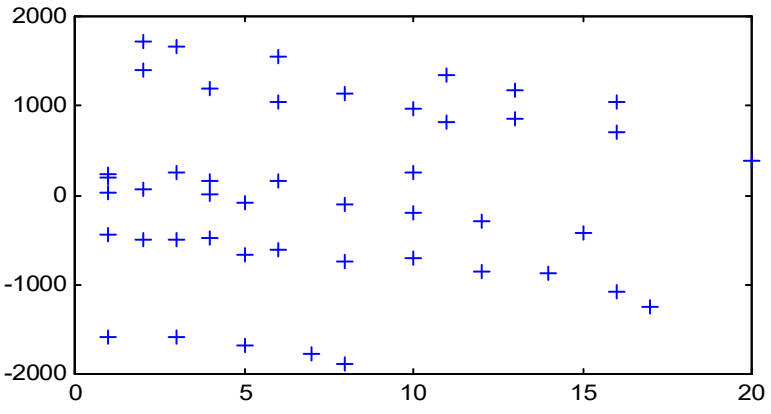
$$\hat{y} = \hat{a}_0 + \hat{a}_1 x_1 + \hat{a}_2 x_2 + \hat{a}_3 x_3 + \hat{a}_4 x_4$$

残差 $e = y - \hat{y}$

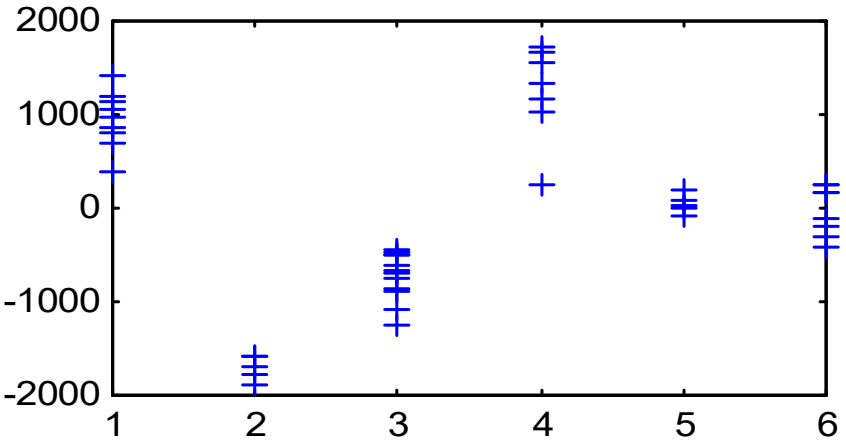
管理与教育的组合

组合	1	2	3	4	5	6
管理	0	1	0	1	0	1
教育	1	1	2	2	3	3

e 与资历x_1的关系



e与管理—教育组合的关系



残差大概分成3个水平，6种管理—教育组合混在一起，未正确反映.

残差全为正, 或全为负, 管理—教育组合处理不当.

应在模型中增加管理x_2与教育x_3, x_4的交互项 .

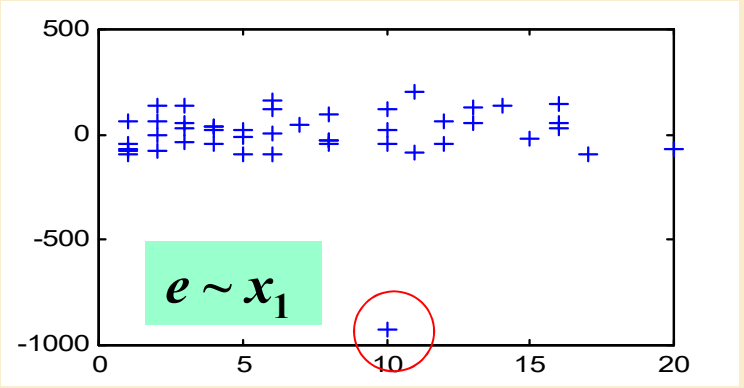
进一步的模型

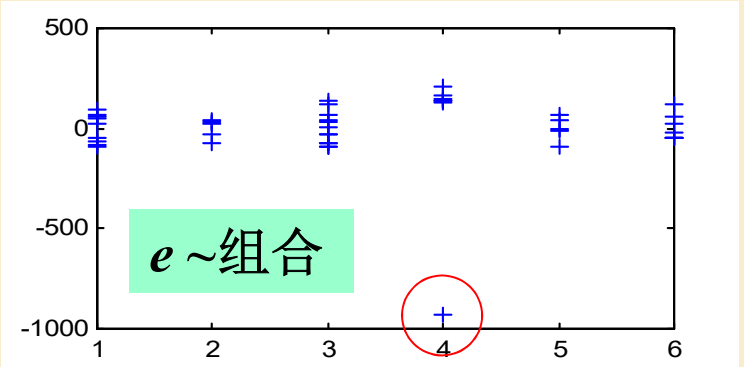
增加管理x_2与教育x_3, x_4的交互项

$$y = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + a_5x_2x_3 + a_6x_2x_4 + \varepsilon$$

参数	参数估计值	置信区间
a_0	11204	[11044 11363]
a_1	497	[486 508]
a_2	7048	[6841 7255]
a_3	-1727	[-1939 -1514]
a_4	-348	[-545 –152]
a_5	-3071	[-3372 -2769]
a_6	1836	[1571 2101]

R^2=0.9988 F=554 p<0.0001 s^2=3×10⁴





R^2, F有改进, 所有回归系数置信区间不含零点, 模型完全可用

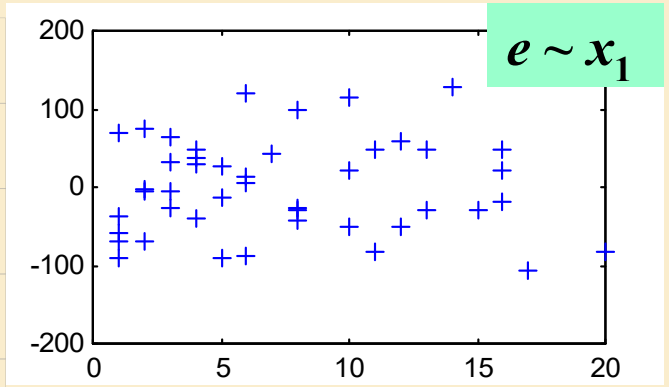
消除了不正常现象

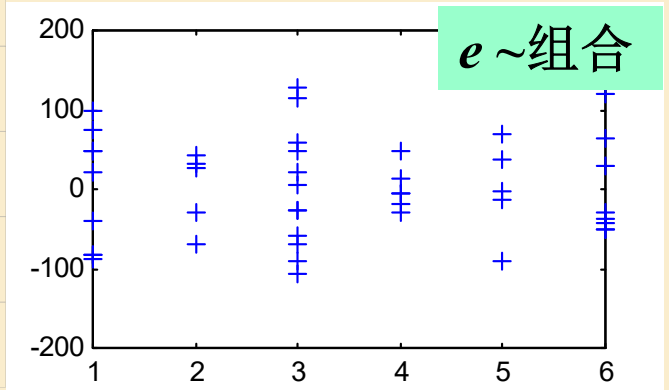
异常数据(33号)应去掉!

去掉异常数据后的结果

参数	参数估计值	置信区间
a_0	11200	[11139 11261]
a_1	498	[494 503]
a_2	7041	[6962 7120]
a_3	-1737	[-1818 -1656]
a_4	-356	[-431 –281]
a_5	-3056	[-3171 –2942]
a_6	1997	[1894 2100]

R^2= 0.9998 F=36701 p<0.0001 s^2=4×10³





R^2: 0.9567 →0.9988→0.9998

F: 226 → 554 → 36701

s^2: 10^4 → 3×10^4 → 4×10^3

置信区间长度更短

残差图十分正常

最终模型的结果可以应用

模型应用

$$\hat{y}=\hat{a}_0+\hat{a}_1x_1+\hat{a}_2x_2+\hat{a}_3x_3+\hat{a}_4x_4+\hat{a}_5x_2x_3+\hat{a}_6x_2x_4$$

制订6种管理—教育组合人员的“基础”薪金(资历为0)

$x_1=0$； $x_2=1$~ 管理，$x_2=0$~ 非管理

中学：$x_3=1, x_4=0$ ； 大学：$x_3=0, x_4=1$； 更高：$x_3=0, x_4=0$

组合	管理	教育	系数	“基础”薪金
1	0	1	a_0+a_3	9463
2	1	1	$a_0+a_2+a_3+a_5$	13448
3	0	2	a_0+a_4	10844
4	1	2	$a_0+a_2+a_4+a_6$	19882
5	0	3	a_0	11200
6	1	3	a_0+a_2	18241

大学程度管理人员比更高程度管理人员的薪金高.

大学程度非管理人员比更高程度非管理人员的薪金略低.

软件开发人员的薪金

对定性因素(如管理、教育)，可以引入0-1变量处理，0-1变量的个数可比定性因素的水平少1.

残差分析方法可以发现模型的缺陷，引入交互作用项常常能够改善模型.

剔除异常数据，有助于得到更好的结果.

注：可以直接对6种管理—教育组合引入5个0-1变量.