10.2

1. 问题：研究软件开发人员的薪金与他们的资历，管理责任，教育程度等因素之间的关系

2. 所有变量都用0,1来进行假设

分析与假设 按照常识，薪金自然随着资历（年）的增长而增加，管理人员的薪金应高于非管理人员，教育程度越高薪金也越高. 薪金记作 y，资历（年）记作 x_1，为了表示是否管理人员，定义

333

$$x_2=\begin{cases}1, & \text{管理人员}\\ 0, & \text{非管理人员}\end{cases}$$

为了表示 3 种教育程度，定义

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

这样，中学用 $x_3=1, x_4=0$ 表示，大学用 $x_3=0, x_4=1$ 表示，研究生则用 $x_3=0$，$x_4=0$ 表示.

为简单起见，我们假定资历（年）对薪金的作用是线性的，即资历每加一年，薪金的增长是常数；管理责任、教育程度、资历诸因素之间没有交互作用，建立线性回归模型.

3. 假设诸因素之间没有相互作用

4. 变量：

- 薪金y
- 资历x1
- 管理责任x2
- 教育程度x3,x4

5. 模型

基本模型 薪金 y 与资历 x_1，管理责任 x_2，教育程度 x_3, x_4 之间的多元线性回归模型为

$$y=a_0+a_1x_1+a_2x_2+a_3x_3+a_4x_4+\varepsilon \qquad (1)$$

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

6. 结果

表 2 模型(1)的计算结果

参数	参数估计值	参数置信区间
a_0	11 032	[10 258，11 807]
a_1	546	[484，608]
a_2	6 883	[6 248，7 517]
a_3	−2 994	[−3 826，−2 162]
a_4	148	[−636，931]

$R^2=0.957$　$F=226$　$p<0.0001$　$s^2=1.057\times10^6$

结果分析 从表 2 知 $R^2=0.957$，即因变量（薪金）的 95.7% 可由模型确定，F 值远远超过 F 检验的临界值，p 远小于 α，因而模型(1)从整体来看是可用的. 比如，利用模型可以估计（或预测）一个大学毕业、有 2 年资历、非管理人员的薪金为

- R^2

判定系数是指可解释的变异占总变异的百分比，用 R^2 表示，有

$$R^2=\frac{SSR}{SST}=(1-\frac{SSE}{SST}) \qquad (15)$$

从判定系数的定义看，R^2 有以下简单性质：

（1）$0\le R^2\le1$；

（2）当 $R^2=1$ 时，有 $SSR=SST$，也就是说，此时原数据的总变异完全可以由拟合值的变异来解释，并且残差为零（$SSE=0$），即拟合点与原数据完全吻合；

（3）当 $R^2=0$ 时，回归方程完全不能解释原数据的总变异，y 的变异完全由与 x

-234-

无关的因素引起，这时 $SSE=SST$。

- F:

$$MSE$$

对于检验水平 α，按自由度（$n_1=1$，$n_2=n-2$）查 F 分布表，得到拒绝域的临界值 $F_\alpha(1,n-2)$。决策规则为

若 $F\le F_\alpha(1,n-2)$，则接受 H_0 假设，这时认为 β_1 显著为零，无法用 x 的线性关系式来解释 y。

若 $F>F_\alpha(1,n-2)$，则否定 H_0，接受 H_1。这时认为 β_1 显著不为零，可以用 x 的线性关系来解释 y。习惯上说，线性回归方程的 F 检验通过了。

需要注意的是，即使 F 检验通过了，也不说明

$$y_i=\beta_0+\beta_1x_i+\varepsilon_i$$

-236-

- 置信区间包含零点

包含零点就是0属于置信区间,这表明估计的系数不显著,举个简单的例子,比如估计出的b = 3.4,但是它的置信区间是[-1,6],则3.4看上去和零查很远,但统计结果是0属于它的置信区间,我们没有显著证据证明真实的b不为0.

F统计量一般是用于对所有系数而言的,它的原假设是所有系数都为0

(参考算法书)

1. 结果分析

进一步的讨论 a_4 的置信区间包含零点，说明基本模型(1)存在缺点. 为寻找改进的方向，常用残差分析方法（残差 e 指薪金的实际值 y 与用模型估计的薪金 $\hat{y}$ 之差，是模型(1)中随机误差 ε 的估计值，这里用了同一个符号）. 我们将影响因素分成资历与管理－教育组合两类，管理－教育组合的定义如表 3.

表 3 管理－教育组合

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

- 组合的原因可能是减少变量的数量

2. 残差分析

为了对残差进行分析，图 1 给出 e 与资历 x_1 的关系，图 2 给出 e 与管理－教育 x_2，x_3，x_4 组合间的关系.

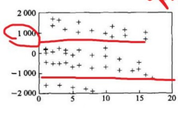

图 1 模型(1)e 与 x_1 的关系

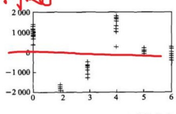

图 2 模型(1)e 与 x_2-x_3,x_4 组合的关系

从图 1 看，残差大概分成 3 个水平，这是由于 6 种管理－教育组合混在一起，在模型中未被正确反映的结果；从图 2 看，对于前 4 个管理－教育组合，残差或者全为正，或者全为负，也表明管理－教育组合在模型中处理不当.

在模型(1)中管理责任和教育程度是分别起作用的，事实上，二者可能起着交互作用，如大学程度的管理人员的薪金会比二者分别的薪金之和高一点.

以上分析提示我们，应在基本模型(1)中增加管理 x_2 与教育 x_3,x_4 的交互项，建立新的回归模型.

更好的模型 增加 x_2 与 x_3,x_4 的交互项后，模型记作

335

图的左下方有两个下拉式菜单，一个菜单Export用以向Matlab工作区传送数据，包括beta(回归系数)，rmse（剩余标准差），residuals(残差)。模型（41）的回归系数和剩余标准差为

beta =−312. 5871　7. 2701　−1. 7337　−0. 0228　0. 0037

rmse =16.6436

一篇关于残差分析的博客

残差

在matlab中 出现1.0e+003 *是什么意思？

匿名 | 浏览 34296 次 | 举报

我有更好的答案

最佳答案　推荐于2017-12-15 12:07:24

您好，

这是科学计数法的表示方式，意思是1*10^3

再举两个例子：

如果您输入了向量[23 000 000, 55 000 000]，那么MATLAB会将之表示为：1.0e+7*[2.3, 5.5]

如果您输入了向量[0.000 000 23, 0.000 000 55]，那么MATLAB会将之表示为：1.0e-7*[2.3, 5.5]