

**《数 学 建 模 B》**

**课程实验报告**

**实验名称：统计回归模型**

**学生班级：信息对抗1602**

**学生姓名：郝希烜**

**班内序号：27**

统计回归模型

**一、实验目的**

（1）着重于数学建模的角度，介绍如何建立若干实际优化问题的模型

（2）在用MATLAB软件求解后，对结果做一些分析.

**二、实验题目**

下表列出了某城市18位3544岁经理的年平均收入（千元），风险偏好度和人寿保险额y（千元）的数据，其中风险偏好度是根据发给每个经理的问卷调查表综合评估得到的，它的数值越大，就越偏爱风险。研究人员想研究此年龄段中的经理所投保的人寿保险额与年均收入及风险偏好度之间的关系。研究者预计，经理的年均收入和人寿保险额之间存在二次关系，并有把握地认为风险偏好度对人寿保险额有线性效应，但对于风险偏好度对人寿保险额是否有二次效应以及两个自变量是否对人寿保险额有交互效应，心中没底。

通过下表中的数据来建立一个合适的回归模型，验证上面的看法，并给出进一步的分析。

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 序号 | y |  |  | 序号 | y |  |  |
| 1 | 196 | 66.290 | 7 | 10 | 49 | 37.408 | 5 |
| 2 | 63 | 40.964 | 5 | 11 | 105 | 54.376 | 2 |
| 3 | 252 | 72.996 | 10 | 12 | 98 | 46.186 | 7 |
| 4 | 84 | 45.010 | 6 | 13 | 77 | 46.130 | 4 |
| 5 | 126 | 57.204 | 4 | 14 | 14 | 30.366 | 3 |
| 6 | 14 | 26.852 | 5 | 15 | 56 | 39.060 | 5 |
| 7 | 49 | 38.122 | 4 | 16 | 245 | 79.380 | 1 |
| 8 | 49 | 35.840 | 6 | 17 | 133 | 52.766 | 8 |
| 9 | 266 | 75.796 | 9 | 18 | 133 | 55.916 | 6 |

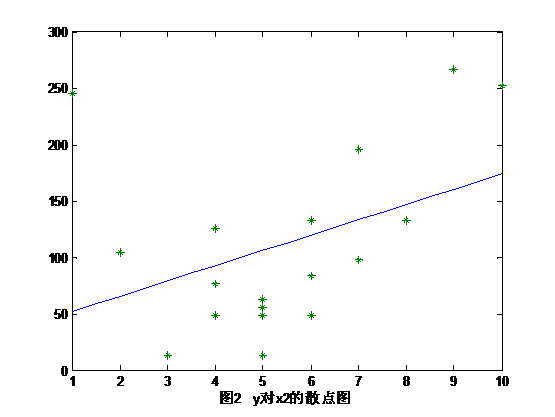
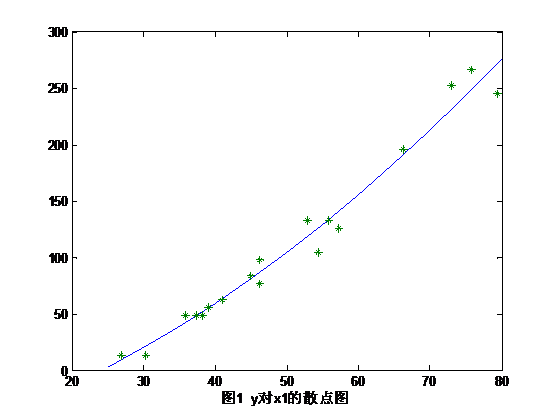
表一 题目数据表

**三、问题分析**

根据我们平常的经验，我们容易做出如下判断：经理的人寿保险额应该随经理人的收入的提升而提高，与该经理人的风险偏好度有着直接的关系（这里是线性关系）。然而，我们并不知道这种关系是二次关系还是线性关系，我们可以通过作图初步判定这种关系。这里我们记人寿保险额y(千元)，年平均收入（千元），风险偏好度.

**四、模型建立**

1.为大致分析y与和关系，首先利用表1的数据分别出y对于和的散点图（见图1和图2中的圆点）



从图1可以发现，随着的增加，y有向上弯曲增加的趋势，图中的曲线是二次函数模型拟合的（其中e是随机误差）

在图2中，当增大时，y的值有比较明显的线性增长趋势，图中的直线式用线性模型拟合的（其中e是随机误差）。

综合上面的分析，结合模型（1）和模型（2）可得如下的回归模型

(3)式右端的和称为回归变量（自变量），是给定年平均收入,风险偏好度 时，人寿保险额y的平均值,其中,,,称为回归系数，有表1的数据估计，影响y的其他因素作用都包含在误差e中，如果模型选择合适，e应大致服从均值为零的正态分布。

2. 交互效应模型：模型（3）中回归变量和对应变量y的影响是相互独立的，即人寿保险金额与年平均收入的二次关系由回归系数确定，而不依赖于风险偏好度。根据直觉和经验可以猜想，和之间的交互作用会对y有影响，不妨简单的用,的乘积代表他们的交互作用，于是将模型（3）增加一项，得到

在这个模型中，y的均值与的二次关系为 ,由系数确定，并依赖于风险偏好度

**五、模型求解**

1. 直接利用matlab统计工具箱中的regress求解

[b,bint,r,rint,stats]=regress(y,x,alpha)

（即附件中1.m）

x=[1 7 66.290 4394.3641

1 5 40.964 1678.0493

1 10 72.996 5328.4160

1 6 45.010 2025.9001

1 4 57.204 3272.2976

1 5 26.852 721.0299

1 4 38.122 1453.2869

1 6 35.840 1284.5056

1 9 75.796 5745.0336

1 5 37.408 1399.3585

1 2 54.376 2956.7494

1 7 46.186 2133.1466

1 4 46.130 2127.9769

1 3 30.366 922.0940

1 5 39.060 1525.6836

1 1 79.380 6301.1844

1 8 52.766 2784.2508

1 6 55.916 3126.5991];

y=[196 63 252 84 126 14 49 49 266 49 105 98 77 14 56 245 133 133]';

[b,bint,r,rint,stats]=regress(y,x,0.05)

其中输入y为模型（3）中y的数据（n维向量，n=18）,x为对应于回归系数a=(a0,a1,a2,a3)的数据矩阵[1 x2 x1 x1^2]，alpha为置信水平@（缺醒时@=0.05）；输出b为a的估计值，计作, bint为b的置信区间，r为残差向量,rint为r的置信区间，stats为回归模型的检验统计量，有3个值，第1个回归方程的决定系数 (R是相关系数)，第2个是F统计量值，第3个是与F统计量对应的概率值p.

matlab计算结果如下：

b =

-62.3489

5.6846

0.8396

0.0371

bint =

-73.5027 -51.1952

5.2604 6.1089

0.3951 1.2840

0.0330 0.0412

r =

-0.0512

0.3076

-1.3718

-0.6730

-3.7605

-1.3560

2.7129

-0.4817

0.5130

-0.3725

0.6842

2.6781

-1.0293

-0.3930

0.5561

1.3578

2.3248

-1.6456

rint =

-3.8062 3.7037

-3.5602 4.1754

-4.4345 1.6909

-4.4953 3.1493

-6.6710 -0.8500

-4.2352 1.5231

-0.7594 6.1852

-4.2421 3.2787

-2.6410 3.6670

-4.2117 3.4667

-2.6689 4.0374

-0.7464 6.1026

-4.7666 2.7080

-3.8380 3.0521

-3.2953 4.4076

-0.4770 3.1926

-1.0603 5.7099

-5.2948 2.0037

stats =

1.0e+04 \*

0.0001 1.1070 0.0000 0.0003

故可以得到模型（3）的回归系数估计值及置信区间（置信水平@=0.05），检验统计量，F，p。

|  |  |  |
| --- | --- | --- |
| 参数 | 参数估计值 | 参数置信度 |
|  | -62.3489 | [-73.5027 -51.1952] |
|  | 5.6846 | [ 5.2604 6.1089] |
|  | 0.8396 | [ 0.3951 1.2840] |
|  | 0.0371 | [0.0330 0.0412] |
| = 1.0e+004 \* 0.0001 F= 1.0e+004 \*1.1070 p=0.0000 | | |

表二 模型（3）计算结果

2.交互效应模型求解

同样利用matlab统计工具箱中的regress求解

[b,bint,r,rint,stats]=regress(y,x,alpha)

（即附件中2.m）

x=[1 7 66.290 4394.3641 464.03

1 5 40.964 1678.0493 204.82

1 10 72.996 5328.4160 729.96

1 6 45.010 2025.9001 270.06

1 4 57.204 3272.2976 228.816

1 5 26.852 721.0299 134.26

1 4 38.122 1453.2869 152.488

1 6 35.840 1284.5056 215.04

1 9 75.796 5745.0336 682.164

1 5 37.408 1399.3585 187.04

1 2 54.376 2956.7494 108.752

1 7 46.186 2133.1466 323.302

1 4 46.130 2127.9769 184.52

1 3 30.366 922.0940 91.098

1 5 39.060 1525.6836 195.3

1 1 79.380 6301.1844 79.380

1 8 52.766 2784.2508 422.128

1 6 55.916 3126.5991 335.496];

y=[196 63 252 84 126 14 49 49 266 49 105 98 77 14 56 245 133 133]';

[b,bint,r,rint,stats]=regress(y,x,0.05)

Matlab计算结果如下：

b =

-65.9461

6.6005

0.8731

0.0374

-0.0138

bint =

-79.6004 -52.2918

4.5786 8.6223

0.4197 1.3265

0.0332 0.0415

-0.0436 0.0160

r =

-0.0092

0.2733

-0.9104

-0.9628

-3.5763

-1.6017

3.0347

-0.9992

1.0091

-0.4486

1.2314

2.1363

-0.7383

0.4176

0.5004

0.5600

1.8146

-1.7311

rint =

-3.8060 3.7877

-3.6387 4.1852

-3.8882 2.0674

-4.7463 2.8207

-6.5133 -0.6392

-4.4002 1.1968

-0.2612 6.3307

-4.5744 2.5761

-1.9248 3.9431

-4.3243 3.4270

-1.8571 4.3199

-1.2282 5.5009

-4.4836 3.0071

-2.5418 3.3771

-3.3950 4.3958

-0.2954 1.4154

-1.5125 5.1417

-5.3910 1.9288

stats =

1.0e+03 \*

0.0010 8.3044 0.0000 0.0033

故可以得到模型（4）的回归系数估计值及置信区间（置信水平@=0.05），检验统计量，F，p。

|  |  |  |
| --- | --- | --- |
| 参数 | 参数估计值 | 参数置信区间 |
|  | -65.9461 | [-79.6004 -52.2918] |
|  | 6.6005 | [4.5786 8.6223] |
|  | 0.8731 | [0.4197 1.3265] |
|  | 0.0374 | [0.0332 0.0415] |
|  | -0.0138 | [-0.0436 0.0160] |
| =1 F=8304.4 p=0.0000 | | |

表三 模型（3）计算结果

表三的结果可知，所有参数的置信区间，和的交互作用项的系数的置信区间包含零点（但右区间距离零点很近），表明回归变量（对应变量y的影响）不是太显著的，但是由于是显著的，我们仍可以将变量保留在模型中。

将回归系数的估计值带入模型（4），即可预测某经理的年平均收入y,预测值记为，得到模型（4）的预测方程：

只需要知道风险偏好度和人寿保险金额,就可以计算预测值.

**六、结果分析与讨论**

**1.结果分析**

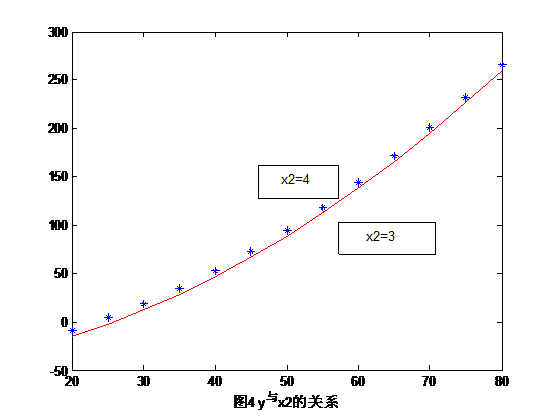
为进一步了解和之间的交互作用，考察模型(4)的预测方程,

如果取风险偏好度=3，带入（5）可得

再取x2=4,带入（5）得：

他们均为的二次曲线图，其图形见图4，且

由式可得，当<343.776时，总有 > ,一般若年均收入相同，风险偏好度高的人，人寿保险金额较高。



通过以上模型（3）和模型（4）的分析和求解过程，可以看到，模型（3）是最理想的，也与我们的假设相一致，则该问题的数学模型为：

只需要知道风险偏好度和人寿保险金额,就可以计算预测值.

**2.模型评价**

此模型的优点是，能通过简单的风险偏好度和年均收入这两组数据得到经理人的人寿保险额,为应用（或使用者）提供了方便。

但是考虑到人寿保险行业的特殊性，影响一个投保人投保额的大小的因素并不只有题中提到的两种，比如投保人的身体健康状况对其投保额的多少就有一定的影响，由于模型只有两个参变量，模型过于粗糙，不能很好地反应现实问题，只能为现实问题提供粗略的估计。应该增加些额外的参变量（当然这些变量应该与保险额相关）对模型加以推广，像职业、健康、年龄这些因素等。