

第十二届“认证杯”数学中国

数学建模网络挑战赛

承 诺 书

我们仔细阅读了第十二届“认证杯”数学中国数学建模网络挑战赛的竞赛规则。

我们完全明白，在竞赛开始后参赛队员不能以任何方式（包括电话、电子邮件、网上咨询等）与队外的任何人（包括指导教师）研究、讨论与赛题有关的问题。

我们知道，抄袭别人的成果是违反竞赛规则的，如果引用别人的成果或其他公开的资料（包括网上查到的资料），必须按照规定的参考文献的表述方式在正文引用处和参考文献中明确列出。

我们郑重承诺，严格遵守竞赛规则，以保证竞赛的公正、公平性。如有违反竞赛规则的行为，我们接受相应处理结果。

我们允许数学中国网站(www.madio.net)公布论文，以供网友之间学习交流，数学中国网站以非商业目的的论文交流不需要提前取得我们的同意。

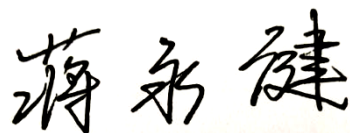
我们的参赛队号为：**3695**

参赛队员（签名）：

队员 1:



队员 2:



队员 3:



参赛队教练员（签名）：

参赛队伍组别：本科组

第十二届“认证杯”数学中国

数学建模网络挑战赛

编 号 专 用 页

参赛队伍的参赛队号：3695

竞赛统一编号（由竞赛组委会送至评委团前编号）：

竞赛评阅编号（由竞赛评委团评阅前进行编号）：

2019 年第十二届“认证杯”数学中国 数学建模网络挑战赛第一阶段论文

题 目 基于广义线性法的车险分析

关 键 词 多元线性分析 广义线性 伽玛分布

摘 要

随着汽车保险业的行业形势加剧，汽车保险定价成为了保险公司中的关键。附件中给我们提供了车险的数据库，要求我们从大数据中分析，在降低成本的同时可以提高客户群的续保率。

问题一中要求我们给出用户的续保的精准画像。给出客户的续保概率。这是一道数据分析的题目，我们首先对于数据进行初步的处理。对于文字数据，我们先进行量化处理。对于离散数据，我们进行聚类分析处理。最后得到了各个因素数据因变量，然后我们进行了多元线性回归拟合，通过线性分析可以逆推筛选出能够影响续保率的因素。公

式为：
$$y = 0.54828X_1 + 0.80737X_2 + 0.0336817X_3 + 0.0405936X_4 + 0.0677043X_6 + 0.004482708X_8 + 0.0135171X_9$$

不过我们还发现模型中存在拟合度不高的情况，因为不能确定各个因素是完全服从线性分布的，以及他们的自相关性，所以我们改进了线性模型，采用广义可加的多元线性模型作为传统线性模型的推广，广义线性模型的核心思想是：用风险分级变量的线性组合的某个函数，来解释损失变量的期望值水平，其函数表达式包括随机成分、系统成分、连接函数。

第二问要求我们以人定价，随着竞争的激烈化和大数据时代的来临，这也要求保险公司能够与时俱进，推陈出新出更个性化的定制方案对于各个用户以增加用户的投保率的同时能够减少自己的成本。所以说我们要 1. 一定的资金回报率 2. 利润最大化 3. 保持或扩展市场份额。所以我们根据线性分析得到了影响因子，制定了价格改革方案：（1）在转移规则的设定中加入索赔额的因素。（2）提高最优折扣级别的折扣比例以及优惠方案的复杂性，以人为车险制定方案，增加影响因子。

参赛队号： 3695

参赛密码 _____
(由组委会填写)

所选题目：_____C_____ 题

Abstract

As the industry situation in the auto insurance industry intensifies, auto insurance pricing has become the key to insurance companies. The attached file provides us with a database of auto insurance, asking us to analyze from big data, and to improve the renewal rate of the customer base while reducing costs.

Question 1 asks us to give an accurate picture of the user's renewal. Give the customer the probability of renewal. This is a topic of data analysis. We first conduct preliminary processing on the data. For text data, we first quantify it. For discrete data, we perform cluster analysis processing. Finally, the data dependent variable of each factor is obtained. Then we carry out multiple linear regression fitting. Through linear analysis, we can reversely select the factors that can affect the renewal rate. The formula is:

$$y = 0.54828X_1 + 0.80737X_2 + 0.0336817X_3 + 0.0405936X_4 \\ + 0.0677043X_6 + 0.004482708X_8 + 0.0135171X_9$$

However, we also found that there is a low degree of fit in the model. Because it is not certain that each factor is completely obeying the linear distribution and their autocorrelation, we have improved the linear model by using a generalized additive multivariate linear model. The generalization of the traditional linear model, the core idea of the generalized linear model is to use a linear combination of risk grading variables to explain the expected value level of the loss variable. Its function expression includes random components, system components, and connection functions.

The second question requires us to price people. With the fierce competition and the advent of the big data era, this also requires insurance companies to keep up with the times and introduce new and more customized solutions for each user to increase the user's insurance rate. At the same time, it can reduce its own cost. So we have to 1. a certain return on capital 2. maximize profits 3. maintain or expand market share. Therefore, we obtained the impact factor based on the linear analysis and formulated the price reform plan: (1) the factor of adding the claim amount in the setting of the transfer rule. (2) Improve the discount ratio of the optimal discount level and the complexity of the preferential scheme, and formulate plans for artificial auto insurance to increase the impact factor.

一. 问题重述

1.1 问题背景

近年，伴随着我国经济以及汽车产业的快速发展，汽车已经进入了我们寻常老百姓家庭，成为必不可缺的一种代步工具。随之而来的道路交通问题，财产理赔问题需要保险产品来使之有所保障。车险产品自推出以来，就成为我国财产保险保费收入的支柱型和龙头险种。

中国目前的车险费率制度，大多数符合“从车主义”。即车险保费多少，主要取决于这辆车本身的各项情况，如车的购置价、座位数、排量、购车年限等。这类方法实现较简单，但是无法反应各个变量之间的准确关系，从而进行合理定价。保险公司的保费收入与其承担的风险不匹配，投保人的投保费和续保率与保险公司的营收不平衡，促使他们更加科学的对车险进行保费定价。

未来的车险定价将逐渐转变为“从人主义”。车险的定价因素直接与驾驶人的驾驶习惯与行驶里程挂钩，通过驾驶人的驾驶行为来判定车险价格，一个具有良好驾驶习惯的车主，可能只需要支付原本保费的 30% 左右，反之驾驶习惯不佳的车主，则会在原本保费的基础上继续上涨。未来中国车险业，同样的一款车，不同的人开，保费价格会完全不同。这个不同可能是取决于投保人本身的驾驶行为，还可能会以投保人本身的年龄、职业、家庭状况等信息为标准。这给了投保人在面对续保的时候有了更多的，更加合理的选择。

保费定价的合理化，必将使得投保人，被投保人真正获益并推动保险行业的持续健康发展。

1.2 需要解决的问题

1. 建立合理的数学模型，对附件中提供的客户进行精准画像，给出客户的续保概率。
2. 针对不同的客户设计不同的优惠和福利方案，以提高续保概率。

二. 问题分析

2.1 问题重要性分析

信息时代的到来，为车险企业提供了一个更加有力的武器，可以通过数字化技术来更加精准地了解客户，制定营销和服务方案。

对投保人来说，有了更多的选择以及更加合理的保费。对保险公司来说，科学合理的保费定价，可以增加投保人的投保率以及续保率，增加保险公司的营收，推动保险行业的健康持续发展。

2.2 问题的思路分析

问题一要求我们对提供的客户进行精准画像，给出客户的续保概率。这是一道数据分析的题目，附件中给了我们大量的数据库要求我们能够建立合适的数学模型去推算客户的续保概率。我们首先要查阅响应的资料先第一步筛选出无关因素以减少计算量，第二部利用数理统计和线性回归的知识建立多元线性回归方程，利用其自相关系数进行逆验证，分析出影响因子。最后在基础上优化我们的模型，建立广义的多线线性回归模型。

第二问要求我们以人定价，随着竞争的激烈化和大数据时代的来临，这也要求保险公司能够与时俱进，推陈出新出更个性化的定制方案对于各个用户以增加用户的投保率的同时能够减少自己的成本。于是我们将个人影响因子筛选出来，分析其对于赔偿率的影响，同样建立多元线性模型，给出定价方案。

2.3 符号说明

符号	意义
ε	随机误差
β	回归系数构成的向量
θ	自然参数
p	离散参数
β_m	第 m 个变量的回归系数
y	续保率
X_i	第 i 个变量的权重值

表 1

三. 模型的假设

5.1 假设的合理性分析

- 1. 假设影响用户续保及保险赔付的原因仅限于附件中所给的条件
- 2. 假设续保率及赔付率服与各项指标从某种可预测的关注
- 3. 假设续保率不受未提取因素的影响
- 4. 假设附件中的数据具有普遍性且真实有效

5.2 模型的可靠性合理性分析

本模型建立时运用大数定理对样本进行分析，并且充分考虑了各个变量直接的相互影响关系。对变量的数值化科学合理，对某些变量进行聚类。研究单个变量对续保率的影响时构建相关函数，验证得出变量对续保率是否存在相关性。最后得出广义的线性回归方程。

5. 附件中的数据真实有效确保了数据的真实性，影响用户续保及保险赔付的原因仅限于附件中所给的条件，避免了无关变量对于续保率的影响；续保率及赔付率服与各项指标从某种可预测的关注，是进行预测的基本条件，并且确保了模型的可建立性。

四. 模型一建立

4.1 变量的选取与度量

4.1.1 大数定理

大数定律又称大数法则、大数率。对一个随机事件，随着试验次数的增加，事件发生的频率趋于一个稳定值;同时，在对物理量测量实践中，大量测定值的算术平均也具有稳定性。在数理统计中，一般有三个定理，贝努力定理和辛钦定理，如:反映算术平均值和频率的稳定性。当 n 很大时，算术平均值接近数学期望;频率以概率收敛于事件的概率。^[1]

对于大数据的处理，我们模型的建立完全依赖于大数定理。我们将在模型中运用大数定理对各个变量进行量化与度量处理。

4.1.2 变量选取

在影响客户续保率的各个因素中，我们选取了渠道，续保年，投保类别，车龄，险种，NCD，被保险人性别，被保险人年龄，立案件数，新车购置价作为主要影响因素作为变量。可以分为从车考虑与从人考虑对续保率进行分析。

从车考虑，车龄，新车购置价毫无疑问会影响客户是否续保，从而影响续保率。这也是车险定价的主要衡量指标之一。

从人考虑，被保险人性别，被保险人年龄，立案件数等因素，作为投保人的投保费用的主要影响因素，投保费用直接与投保人的续保率挂钩。^[2]

4.1.3 变量的度量与数值化

我们将保费的购置渠道，对某一个变量数值化时，选取当其他量续保率最大时的值，作为其他变量的固定值，然后计算得出此变量变化时的续保率。可得：

渠道	续保率
电网销	0.21194605
个人代理	0.150458716
交叉销售	0.073333333
车商渠道	0.359536082
普通兼代(含银行代理)	0.022727273
专业中介	0.069767442

图 1

4.2 变量的聚类处理

聚类是将相同或相似的对象放在同一个类中，不相似的对象放在不同类中。使得差异趋向两极化，即同类间差异小，不同类间差异大。聚类分析的技术到目前为止是数据挖掘领域发展最为成熟的技术之聚类分析方法主要包括统计学方法和机器学习的方法。作为统计学的一个分支,人们很早就已经对聚类分析有了比较深入的研究。在统计学中,聚类一般称为聚类分析,主要研究基于几何距离的聚类,它首先要定义多维空间和距离,计算权重,以距离作为相似性的判别标准。在机器学习中,聚类称为无监督学习 (Unsupervised Study),它主要是一种不依靠事先确定数据类别,或标有数据类别的学习训练样本集合的学习过程,即它是一种观察学习过程,而不是实例学习过程,它与分类是相对应的。

图 2 展示了广泛采用的有反馈调整的聚类分析模型，主要包括：

第一步：数据样本进行预处理，如特征选取，数据标准化等；

第二步：对样本数据进行聚类，包括聚类算法选择，聚类的参数设置等；

第三步：对上一步得到的聚类进行有效性评价，最终得到聚类结果；

4.2.1 层次聚类方法

层次聚类方法也是发展比较早、应用比较广泛的一大类聚类分析方法，该方法通过将数据组织成若干组，并形成一种树形的聚类结构。聚结型层次聚类法最开始将每一个对象作为一个子类，然后将这些子类按照最近原理进行合并以构造逐渐增大的类别，最后形成一个大类，通常是满足一定条件为止。

在层次聚类方法中，各个类之间的相似程度通常用距离来表示，下面四种距离比较广泛地用于计算两个类之间的差异度：

①平均值距离： $d(C_i, C_j) = d(f_i, f_j)$

②平均距离： $d(C_i, C_j) = \frac{1}{n_i n_j} \sum_{p^{(i)} \in C_i} \sum_{p^{(j)} \in C_j} d(p, p)$

③最大距离: $d(C_i, C_j) = \max_{p^{(i)} \in C_i, p^{(j)} \in C_j} d(p^{(i)}, p^{(j)})$

④最小距离: $d(C_i, C_j) = \min_{p^{(i)} \in C_i, p^{(j)} \in C_j} d(p^{(i)}, p^{(j)})$

C_i, C_j 是聚结过程中同一层次上的两个类, n_i 和 n_j 分别是 C_i 和 C_j 两个类中的对象数目, $p^{(j)}$ 为 C_j 中的任意一个对象, $p^{(i)}$ 为 C_i 中的任意一个对象, f_i 为 C_i 中对象的平均值, f_j 为 C_j 中对象的平均值。^[3]

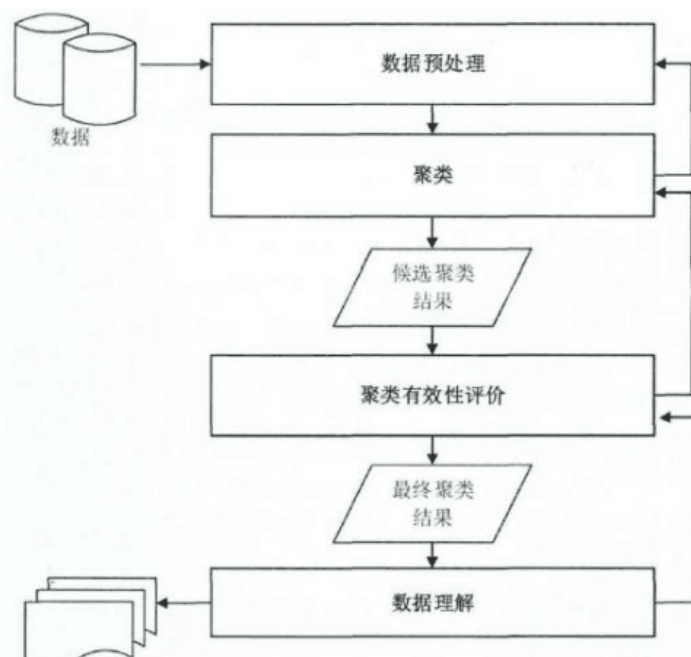


图 2

比较传统的层次聚类基本算法有 AGNE (Agglomerative NESTing) 算法和 DIANA (Dicisivi Analysis) 算法, 分别为聚结型层次聚类算法和分解型层次聚类算法。传统的层次聚类算法比较简洁, 易于理解和应用, 但是聚类的结果受各个类的大小和其中对象分布形状的影响, 适用于类的大小相似且对象分布为球形的聚类。在对象分布形状比较特殊的情况下, 可能会产生错误的聚类结果。另外, 每一次类的聚结或分解都是不可逆的, 并直接影响着下一步的聚结或分解。如果某一步的聚结或分解不理想, 形成聚类的质量就可能很低。

近几年, 出现了一些新的属于层次聚类方法的聚类算法, 一般采用的是聚结型层次聚类策略, 例如: BIRCH (Balanced Iterative Reduring and Clustering using Hierarchies) 算法 CURE (Clustering Using Representatives) 算法、ROCK 算法和 Chameleon 算法。

①BIRCH 算法扫描数据库建立一个初始基于内存的特征树, 该树被晋作是对数据的压缩, 并且包含着数据中所包含的有关聚类结构的内涵。

②BIRCH 算法应用一个聚类算法对特征树的叶子进行聚类。BIRCH 算法通过了聚类特征以及聚类特征树的概念对数据压缩, 对数据集进行一次扫描就可以形成质量较好的聚类, 并且可以通过追加扫描进一步提高聚类的质量。因此, 该算法适用于大规模数据

的聚类。但是，该算法采用了半径或直径的概念来限制类的分布范围，所以适用于对象分布为球形的情况。除此之外，该算法受数据输入顺序的影响。^[3]

4.2.2 变量聚类

对于数据中的某一些变量，例如：投保人年龄，签单保费等；微小的变化对于续保率没有明显的影响。比如，34 岁的客户以及 35 岁的客户，在其余变量相同时，续保率大致相等。同理对于签单保费，根据签单保费的分布，对其进行聚类，结果如图所示：

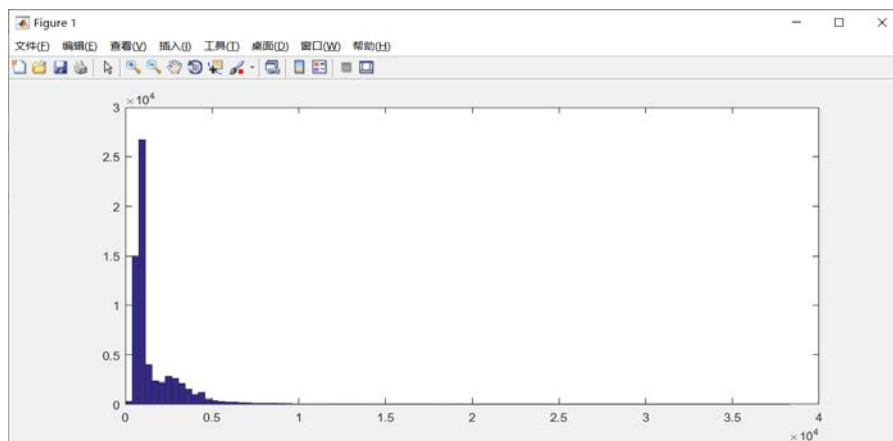


图 3 签单保费分布

签单保费	总数量	续保数	续保率
500	38054	7175	0.18854785
1500	10740	1357	0.12635009
2500	6859	1517	0.22116927
3500	4556	1061	0.23287972
4500	2242	453	0.20205174
5500	3084	1258	0.4079118

图 4 签单保费与续保率

4.3 多元线性回归模型

4.3.1 多元线性回归模型

多元线性回归模型的一般形式为：

$$\eta(u) = \beta_1 \varphi_1(u) + \beta_2 \varphi_2(u) + \cdots \beta_m \varphi_m(u) \quad (3-1)$$

$$\text{令 } y = \beta_1 \varphi_1(u) + \beta_2 \varphi_2(u) + \cdots \beta_m \varphi_m(u) + \varepsilon \quad (3-2)$$

其中， ε 为随机误差，且服从 $N(0, \sigma^2)$ ， $\varphi_i(u)(i=1,2,\cdots,m)$ 均为实际问题的解释变量，是已知函数。

$$\begin{matrix} u_1 & y_1 \\ \vdots & \vdots \\ u_n & y_n \end{matrix}$$

假设做了 n 次试验，得到 n 组的观测值为：

$$\begin{cases} y_i = \beta_1 \varphi_1(u_1) + \beta_2 \varphi_2(u_2) + \cdots + \beta_m \varphi_m(u_m) + \varepsilon_i \\ i=1, 2, \dots, n \end{cases} \quad (3-3)$$

其中, ε_i 为第 i 次试验的随机误差, 且相互独立同服从于 $N(0, \sigma^2)$ 。

该模型关于回归系数 $\beta_1, \beta_2, \dots, \beta_m$ 为线性的, u 一般是向量。为了方便, 引入矩阵记

$$\text{号: } Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} \varphi_1(u_1) & \varphi_2(u_1) & \varphi_3(u_1) & \varphi_4(u_1) \\ \varphi_1(u_2) & \varphi_2(u_2) & \varphi_3(u_2) & \varphi_4(u_2) \\ \vdots & \vdots & \vdots & \vdots \\ \varphi_1(u_n) & \varphi_2(u_n) & \varphi_3(u_n) & \varphi_4(u_n) \end{bmatrix}, \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix}, \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

其中 X 称为模型设计矩阵, 是常数矩阵, Y 与 ε 是随机向量, 且:

$$Y \sim N_n(X\beta, \sigma^2 I), \varepsilon \sim N_n(0, \sigma^2 I) \quad , \quad (I \text{ 为 } n \text{ 阶单位阵}), \quad \varepsilon \text{ 是不可观测的随机误差向量, } \beta \text{ 是回归系数构成的向量, 是未知待定的常数向量。}^{[4]}$$

4.3.2 多元线性模型的理论与改进

多元线性回归模型在非寿险分类费率的厘定中有很广泛的应用, 但其严格的假设条件在非寿险中通常难以得到满足;

首先, 要求因变量服从正态分布在很多情况下可能不行, 续保率等通常不会服从正态分布。

其次, 非寿险的因变量如索赔频率和次均赔款等通常是非负的, 而正态分布的假设显然不能满足这一要求。

第三, 如果因变量是严格非负的, 那么从直观上看, 当因变量的均值趋于零时, 其方差也应该趋于零, 即因变量的方差应该是其均值的函数。但在多元线性回归模型中, 假设因变量的方差是固定的常数, 与均值没有任何关系。

第四, 在多元线性回归模型中, 假设费率因子通过加法关系对因变量产生影响, 但在很多情况下, 费率因子之间可能是一种乘法关系, 而非加法关系。

4.4 广义线性模型

广义线性模型由 Nelder 和 Wedderburn 提出, 是常见的正态线性模型的推广形式, 相比正态线性模型, 广义线性模型保持其框架不变, 但自变量对响应变量函数的影响表现为线性形式, 对其推广主要表现在下面两个方面:

(1) 在因变量分布方面, 广义线性模型扩充了分布类型。正态线性模型假设因变量服从或者近似服从正态分布, 而广义线性模型假设因变量 Y 服从指数分布族, 分布族中包括的常见分布有: 正态分布、二项分布、泊松分布、伽玛分布和逆高斯分布等。其指数分布族的密度函数可以表示为:

$$f(y|\theta, \phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\} \quad (3-4)$$

其中, a, b, c 为已知函数, 对所有的观测值具有相同的形式。 θ 为自然参数, 与均值是一一对应关系, p 为离散参数。通过下面简单的计算可得到指数分布族的均值和方差。

$$\text{由} \quad \int \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\} dy = 1 \quad (3-5)$$

对积分号两边求导 $\int \frac{y-b'(\theta)}{a(\phi)} \exp\left\{\frac{y\theta-b(\theta)}{a(\phi)} + c(y, \phi)\right\} dy = 0$

$$\text{即} \quad \mu = E(Y) = b'(\theta) = \frac{\partial b(\theta)}{\partial \theta} \quad (3-6)$$

二次求导，整理得

$$\int y^2 \exp\left\{\frac{y\theta-b(\theta)}{a(\phi)} + c(y, \phi)\right\} dy - \mu^2 = a(\phi) \bullet b''(\theta)$$

$$\text{Var}(Y) = a(\phi) \bullet b''(\theta) = a(\phi) \bullet \frac{\partial^2 b(\theta)}{\partial \theta^2}$$

变换后，在表达形式方面，广义线性模型是假设响应变量均值经过某联结函数等于解释变量的线性组合形式，具体如下：

$$\mu_i = g^{-1}(\eta_i) = g^{-1}(x_i' \beta_i) \quad (3-7)$$

其中， x_i 是解释向量，常用的联结函数有对数联结函数、倒数联结函数和 logit 联结函数等。对于响应变量服从不同的分布其联结函数与之对应，譬如对于泊松分布选择对数联结函数，伽玛分布选择倒数联结函数，二项分布假设下则常选择 logit 联结函数等等。

模型中参数可利用极大似然估计法来估计，通过加权最小二乘估计，再使用 Newton-Raphson 迭代算法，可得到八的极大似然估计，也可以通过贝叶斯估计法，利用蒙特卡罗模拟得到估计值。因为计算过于复杂，可通过 R, SAS 和 BayesX 等软件进行拟合估计。

4.5 广义可加模型

广义可加模型是在广义线性模型的基础框架下，将解释变量的形式引入非参数函数，使得模型更一般化，从而具有较小的偏差、良好的稳健性和广泛适用性等优点。具体来看，广义可加模型由三个部分组成：随机部分、系统部分和联结函数。

第一部分是随机部分，是指响应变量 Y 的概率分布。与广义线性模型一样，假设响应变量的每个观测值相互独立且服从指数分布族，如正态分布、泊松分布、伽玛分布、逆高斯分布和 Tweedie 分布等。

第二部分是系统部分，与广义线性模型不同的是广义可加模型加入了非参数函数部分，使得适用范围更加广阔，尤其对某些对因变量影响较大，却不好划分等级的变量，可以说是找到了解决的方法。具体表达形式为：

$$\eta_i = x_i' \beta_i + \sum_{k=1}^r f_k(x_k) \quad (3-8)$$

第三部分是联结函数，通过联结函数将响应变量与解释变量连接一起，其表达形式为：

$$g(\mu_i) = \eta_i \quad (3-9)$$

广义可加模型参数估计常用的方法有惩罚最小二乘法，这种方法简单实用，目前研究应用最广泛，但缺陷是忽略对样本先验信息的利用。而贝叶斯方法是提出估计广义可加模型参数的新方法，其充分挖掘参数的先验信息，吸引了广大学者的研究。^[5]

4.6 模型一

根据提供的数据，我们选取了 10 种影响因素作为变量，利用大数定理，变量的数值化以及聚类处理重新建立多元线性回归模型。我们可以得到以下数据：

	渠道	续保年段	投保类别	车龄段	险种
Coedd	0.54828	0.80737	0.0336817	0.0405936	0
t-stat	16.63391	57.5059	-8.2743	4.9970	1.4641
p-val	0.0000	0.0000	0,000	0.0000	0.1432
	NCD 赋值	被保险人性别	年龄段	立案件数	新车购置价
Coedd	0.0677043	0	0.00482708	0.0135171	0
t-stat	3.9227	0.9426	-2.1675	3.7947	1.3408
p-val	0.0001	0.3459	0.0302	0.0001	0.1800

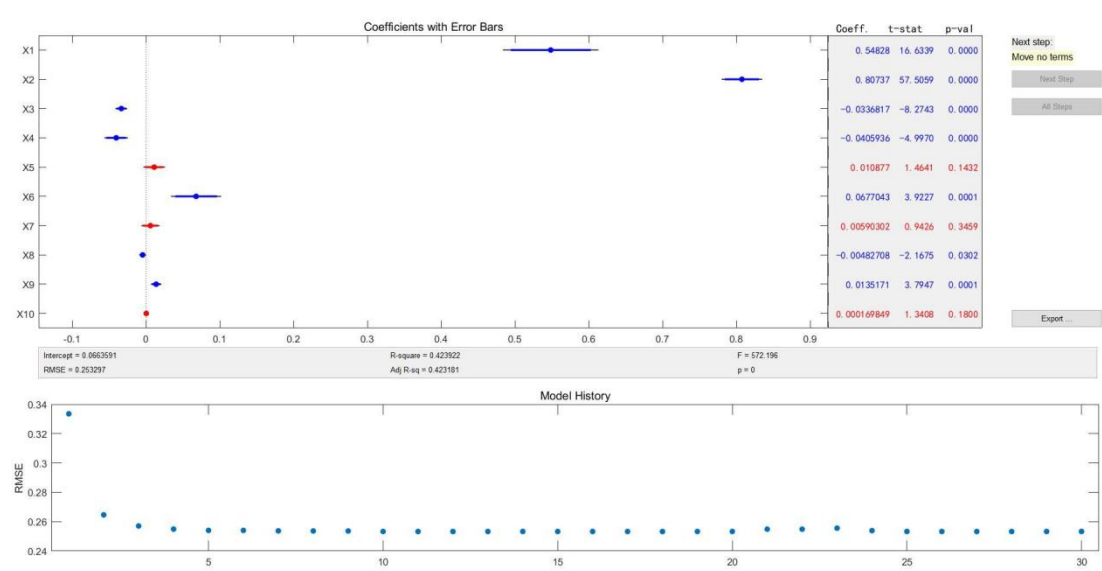


图 5

得到线性回归方程：

$$y = 0.54828X_1 + 0.80737X_2 + 0.0336817X_3 + 0.0405936X_4 + 0.0677043X_6 + 0.004482708X_8 + 0.0135171X_9 \tag{3-10}$$

由图 5 可得 R 值为 0.423 说明回归方程对样本的拟合度相对较好。

我们对每一个回归系数作 p-val。对于每一个因素的 p-val，其值越大，表明对结果的影响越不显著。因此，我们将 p-val 大于 0.00005 的系数置为 0，表明其对续保率没有显著的影响。

利用 MATLAB 编程，我们可以得到回归模型的残差图：

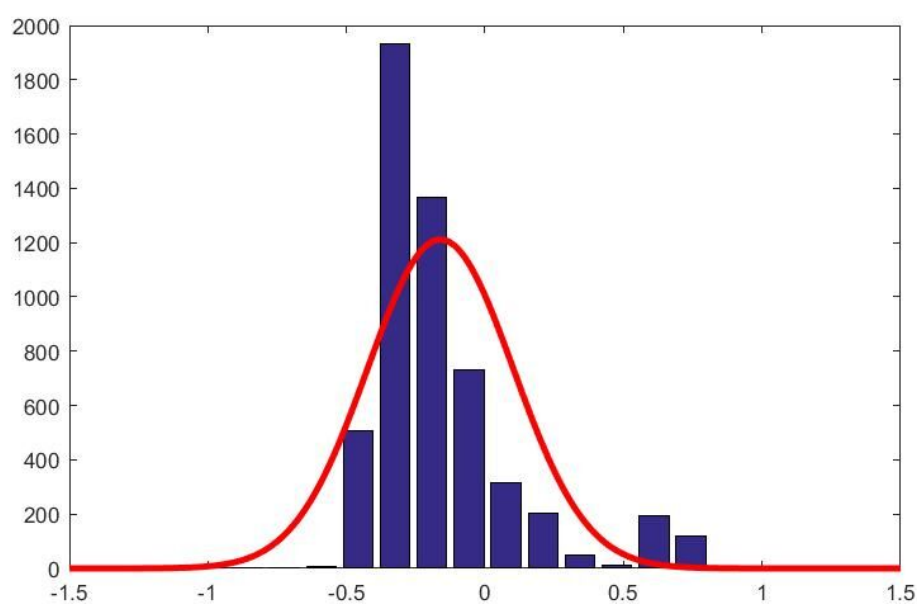


图 6 残差图

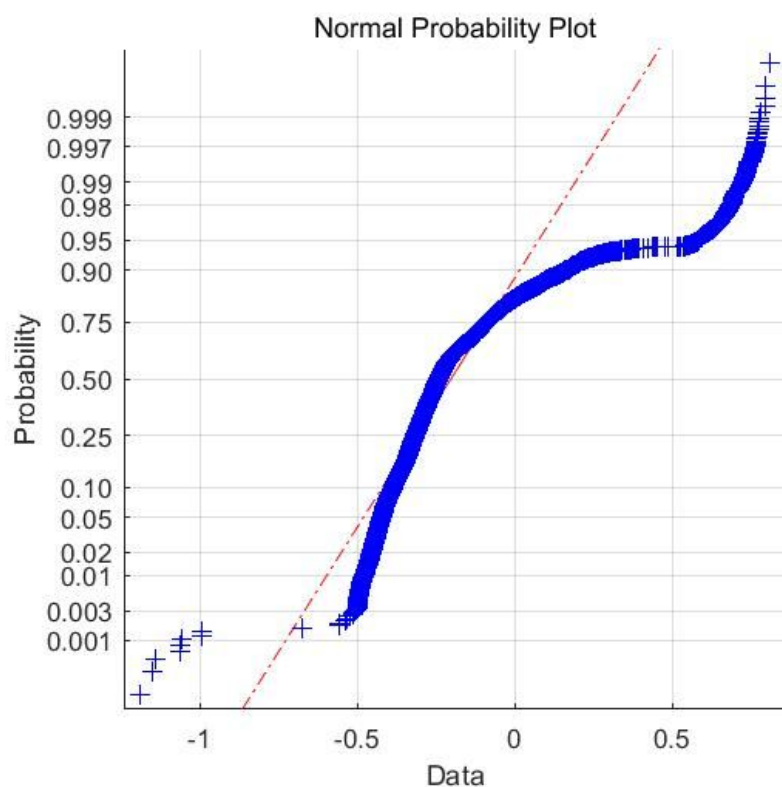


图 7 正态分布检验

可以看到，残差图基本上服从正态分布，误差项除个别点外，基本上都分布在 45 度线附近，所以残差通过正态性检验，满足基本假设。

由于异方差的存在使最小二乘估计量不再是最好的无偏估计量，会导致模型的残差不再是同方差正态分布的，所以要进行异方差检验，我们可以得到残差散点图：

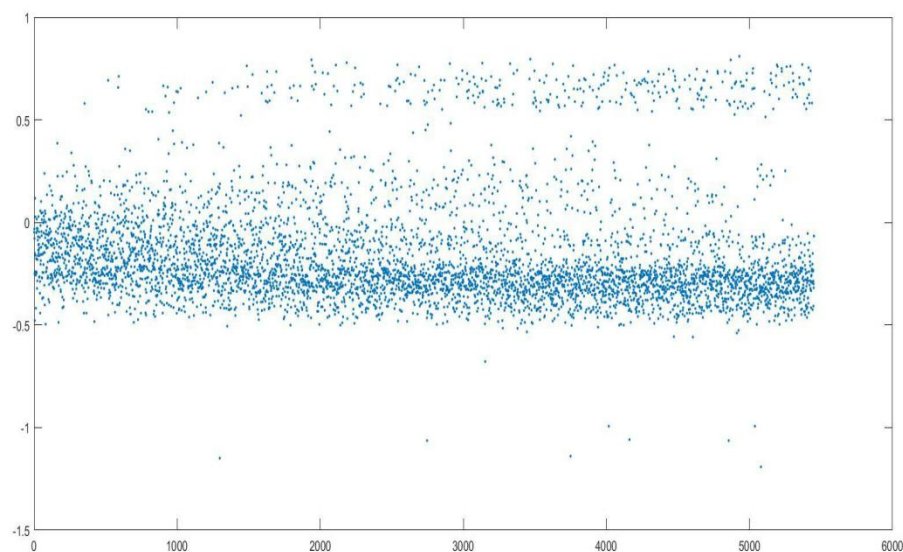


图 8 残差散点图

这里我们可以看出残差的散点图比较分散，所以我们得出的回归模型不存在异方差性。

五. 模型二建立

4.1 问题分析

保险产品定价可以大致分为两个步骤:第一步是建立充分费率，第二步是设定其实际价格。充分费率是在持续经营的假设条件下，满足保险公司的长期利润目标的费率。充分费率的厘定应遵循充足性、公平性、合理性、可行性、稳定性和弹性原则。简而言之，就是使投保人承担的风险与其缴纳的保费相匹配。

保险定价是在充分费率的基础上，结合市场分析设定保险价格。保险定价一般要综合考虑以下三个目标：

1. 一定的资金回报率
2. 利润最大化
3. 保持或扩展市场份额

汽车交通是由人、车、环境组成的一个复杂的系统，它具有很强的不稳定性。它的安全性取决于不固定的人的判断与行动，因而必须把人这一要素作为基本出发点。同时，还要改善车辆与环境，以此来减少事故发生的诱因以及减轻事故灾害的程度。

可以看出我国的NCD系统还是存在很多问题的。在根据我们的数学模型建模分析后，我们提出以下几点改变：

1. 折扣等级、转移规则太简单粗糙，不能对被保险人的承担风险水平进行有效的细分。
2. 惩罚过于温和。
3. 折扣的计算基础是基础保费，即按照保险监管部门批准的费率规章计算出的保险费。

我国汽车保险的NCD制度对占绝大多数的低风险投保人所收取的保费过高，对少数的高风险投保人的惩罚过低，以及对风险分级的简单粗糙，使其无法对风险作出比较严

格的细分，使保险公司失去了财务上的稳定性，并且造成对多数投保人的不公平。长久下去，将对我国车险市场健康发展造成严重的不利影响。
针对以上状况，可以考虑从以下几方面进行改善：

(1) 增加折扣级别数。

折扣级别数直接关系到 NCD 在减少风险不均匀性方面的作用。级别数越多，NCD 的作用越有效。理论上讲，不断增加的折扣组别可以完全消除风险的不均匀性(即非同质性)，使得保费分配更加公平。

(2) 提高最优折扣级别的折扣比例。

折扣比例也是各国监管机关用来改善异质风险的常用工具。折扣比例越大，就越能拉开不同驾驶记录驾驶员的保费差距。在 10070-30%的低折扣比例下，出险率高的驾驶员并不比出险率低的驾驶员多支付多少保费，这对驾驶员就形不成强激励，那么无赔款优待的初衷就难以实现。

(3) 在转移规则的设定中加入索赔额的因素。

我国大多数保险公司的奖惩系统在根据驾驶员的驾驶记录来确定其级别时，通常只考察了他的索赔次数，对于其索赔金额并不计较。这不仅对索赔金额小的投保人不公平，也不利于保险人对投保人进行风险评估。^[6]

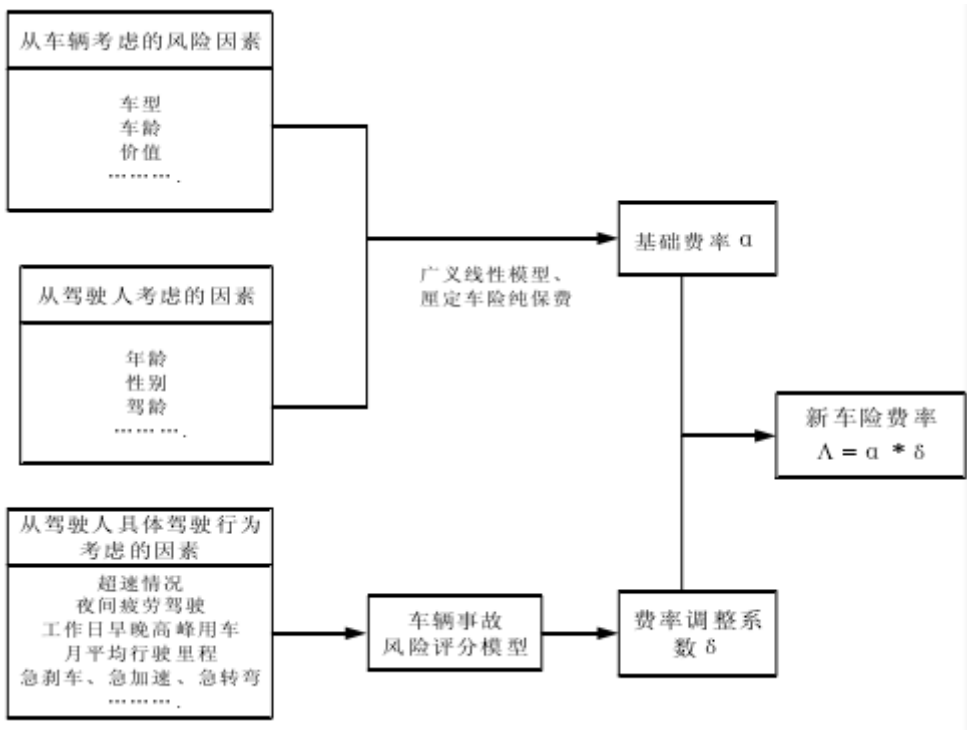


图 9

5.2 广义线性回归模型

我们选取：1 续保年段 2 投保类别 3 车龄段 4 险种 5 NCD 赋值 6 被保险人性别 7 年龄段 8 立案件数 9 新车购置价；9 个因素作为影响“赔付率”的变量。而后根据数据内容，对这 9 个变量进行广义线性回归，可得图 10：

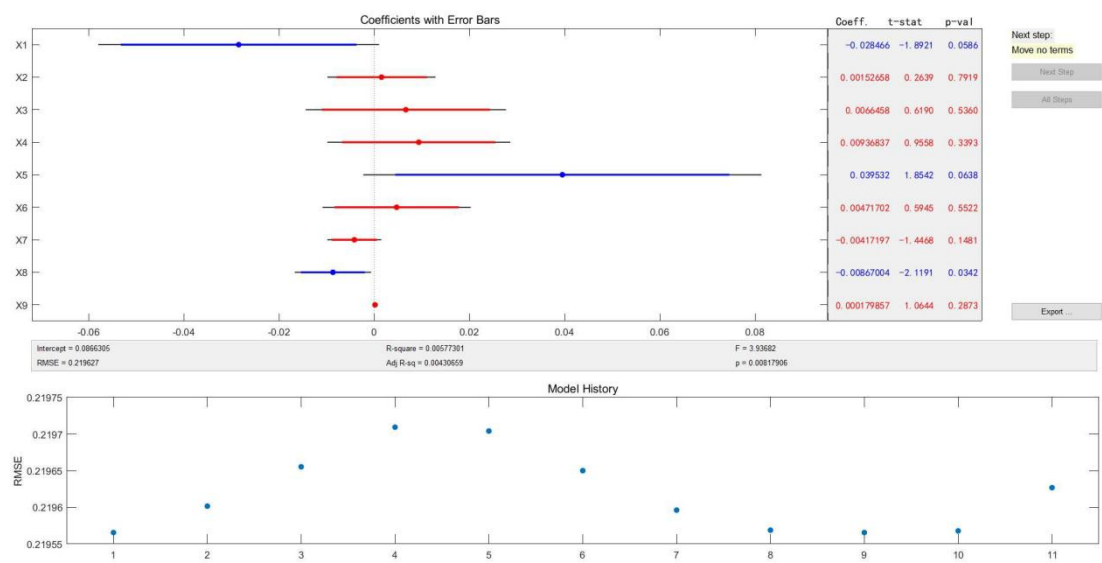


图 10

由图可见：

(1) 续保年段与赔付率成负相关，意味着续保年段的增大，会致使赔付率降低，从而可以得出投保人的驾驶行为以及车辆情况较为良好，这时应当适当的对投保金额进行折扣，从而增大续保概率。

(2) 从驾驶人的性别来看，驾驶人的性别差异对交通事故有着显著的影响，从而对赔付率有着显著的影响。从男女性的角度来讲，重大事故男性驾驶人居多，而小的擦碰则为女性驾驶人居多。所以对不同性别的投保人提出不同的续保优惠方案，可以有效的提高续保率。例如：相同相同签单保费情况下，提高女性投保人的重大事故险保额，适当降低其擦挂事故的险保额。

5.3 改进的广义线性回归模型

对于因素中的几个非线性因素，我们对其进行推广为一元函数。对非线性因素单独做 h_v 函数，观察其独立的函数分布。如图所示（以年龄段为例）：

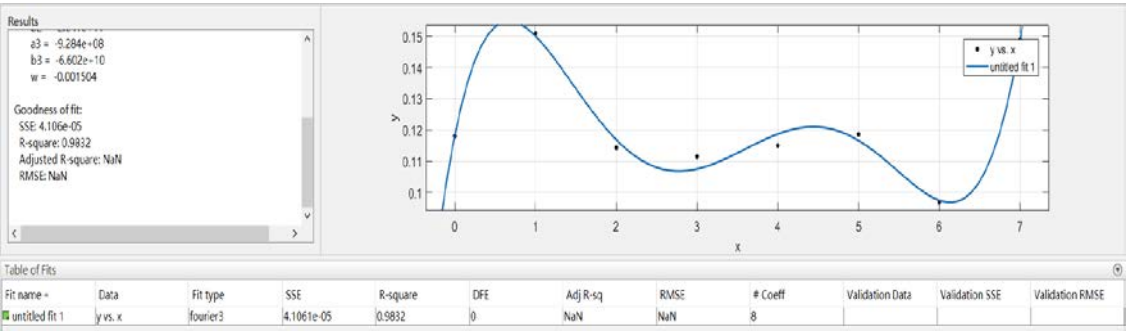


图 11

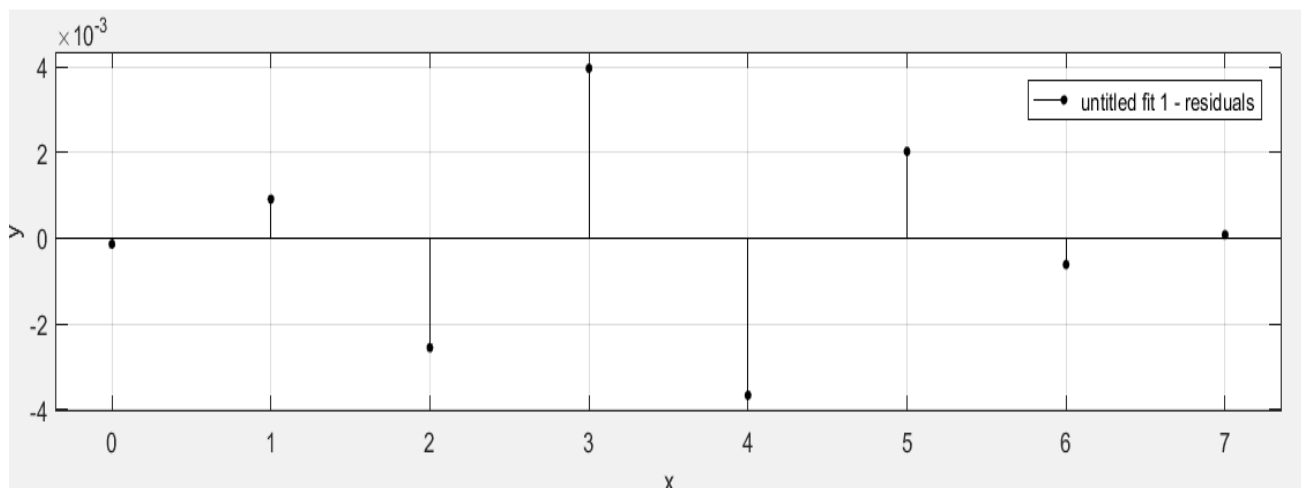


图 12

我们可以看出对于驾驶人的年龄，中年驾驶人较之年轻驾驶人，更加稳重且富有经验，所以事故发生率相对较低。年轻驾驶人相对意气用事，经验不足，发生事故概率较高。

所以，针对这 9 个因素的广义线性回归方程，我们可以得出各个投保人按照 9 个因素的权值比重后得出的续保率与赔付率，从而对其定制合理的优惠以及福利方案，从而提高续保率。

六. 模型的评价与推广

6.1 模型的优缺点

模型的优点：

利用聚类分析和量化减少了数据分析量。在利用线性分析的时候理由显著性检验筛选出符合线性变换的因素。在之后引入了广义线性变换丰富了原因变量，也使得模型更加的完备。

模型的缺点：

利用广义线性性的时候因为要判断出各种函数模拟加大了数据处理量。最后的结果不够量化清晰。

6.2 模型的推广

广义可加的线性数学模型可以在很多线性分析上用到，之前传统的多元线性回归只能模拟出一次函数的关系，有很多多元分析问题多因变量往往因为不是遵从一次函数的关系导致拟合失败。所以在处理这一部分问题的时候就可以先进行广义的函数拟合分析，去处理这一类问题。

参考文献

- 问
- 年
- [1] 360 百科 大数定理: <https://baike.so.com/doc/6681488-6895385.html> 2019.4.20 访问
 - [2] 胡伊 《基于多源数据的车辆风险分析与保险定价模型研究》 64-65 页 2016 年
 - [3] 杨明瀚 《分类变量数据聚类内部评价及算法研究》 7-19 页 2018 年
 - [4] 韩中庚 《数学建模方法及其应用》第二版 高等教育出版社 2009 年
 - [5] 刘忠义 《基于广义可加模型的汽车保险定价研究》 19-30 页 2017 年
 - [6] 许家天 《无赔款优待系统的应用与改进》 20-59 页 2008 年

附录

相关代码 (MATLAB) :

赔付率计算:

```

clc;
statistical= xlsread('统计-续保率.xlsx');
statistical(isnan(statistical)) = 0;
NUM_COUNTER=zeros(size(statistical,1),1);
STATISTIC_COUNTER=zeros(610,13);
parameter=[1 3 5 7 9 11 13 15 16 19 20];
k=1;
for i=1:size(statistical,1)
    if NUM_COUNTER(i)==0 %确认当前行是否被统计
        NUM_COUNTER(i)=1;
        STATISTIC_COUNTER(k,1)=statistical(i,parameter(1));
        STATISTIC_COUNTER(k,2)=statistical(i,parameter(2));
        STATISTIC_COUNTER(k,3)=statistical(i,parameter(3));
        STATISTIC_COUNTER(k,4)=statistical(i,parameter(4));
        STATISTIC_COUNTER(k,5)=statistical(i,parameter(5));
        STATISTIC_COUNTER(k,6)=statistical(i,parameter(6));
        STATISTIC_COUNTER(k,7)=statistical(i,parameter(7));
        STATISTIC_COUNTER(k,8)=statistical(i,parameter(8));
        STATISTIC_COUNTER(k,9)=statistical(i,parameter(9));
        STATISTIC_COUNTER(k,10)=statistical(i,parameter(10));
        for j=1:size(statistical,1) %统计数量
            if
STATISTIC_COUNTER(k,1)==statistical(j,parameter(1))&&STATISTIC_COUNTER(k,2)
==statistical(j,parameter(2))&&STATISTIC_COUNTER(k,3)==statistical(j,parame
ter(3))&&STATISTIC_COUNTER(k,4)==statistical(j,parameter(4))&&STATISTIC_COU
NTER(k,5)==statistical(j,parameter(5))&&STATISTIC_COUNTER(k,6)==statistical
(j,parameter(6))&&STATISTIC_COUNTER(k,7)==statistical(j,parameter(7))&&STAT
ISTIC_COUNTER(k,8)==statistical(j,parameter(8))&&STATISTIC_COUNTER(k,9)==st
atistical(j,parameter(9))&&STATISTIC_COUNTER(k,10)==statistical(j,parameter
(10))
                STATISTIC_COUNTER(k,11)=STATISTIC_COUNTER(k,11)+1;
                NUM_COUNTER(j)=1;
                if statistical(j,parameter(11))==1
                    STATISTIC_COUNTER(k,12)=STATISTIC_COUNTER(k,12)+1;
                end
            end
        end
        STATISTIC_COUNTER(k,13)=STATISTIC_COUNTER(k,12)/STATISTIC_COUNTER(k,11)
;
        k=k+1;
    end
end

```

```

        end
    end
    x=zeros(size(STATISTIC_COUNTER,1),size(parameter,1));
    for i=1:size(parameter,2)-1
        x(:,i)=STATISTIC_COUNTER(:,i);
    end
    y=STATISTIC_COUNTER(:,13);
    stepwise()
        beta=[0.54828      0.80737      0.0336817      0.0405936      0 0.0677043      0
0.00482708 0.0135171      0]
        beta_0=0.0663591;
        z=zeros(size(y,1),1)
        for i=1:size(x,1)
            for j=1:10
                z(i)=z(i)+beta(j)*x(i,j);
            end
            z(i)=z(i)+beta_0;
        end
        delta=zeros(size(y,1),1)
        for i=1:size(y,1)
            delta(i)=y(i)-z(i);
        end
        histfit(delta);
        normplot(delta);
        [y2 x2]=hist(delta,15);
        bar(x2,y2);
        hold on;
        [muhat,sigmahat,muci,sigmaci]=normfit(delta);
        x1 = -1.5:0.0001:1.5;
        y1= normpdf(x1, -0.1609, 0.2635)*800;
        plot(x1,y1,'r','LineWidth',3);

```

计算签单保费与续保率关系时代码:

```

clc;
premium= xlsread('签单保费.xlsx');
premium(isnan(premium)) = 0;
for i=1:size(premium,1)
    if premium(i,1)<=1000
        premium(i,2)=500;
    elseif premium(i,1)>1000 && premium(i,1)<=2000
        premium(i,2)=1500;
    elseif premium(i,1)>2000 && premium(i,1)<=3000
        premium(i,2)=2500;
    elseif premium(i,1)>3000 && premium(i,1)<=4000

```

```

        premium(i, 2)=3500;
    elseif premium(i, 1)>4000 && premium(i, 1)<=5000
        premium(i, 2)=4500;
    else premium(i, 1)>5000
        premium(i, 2)=5500;
    end
end
%统计开始
STAT=zeros(6, 4);
STAT(:, 1)=[500, 1500, 2500, 3500, 4500, 5500];
for k=1:6
    for i=1:size(premium, 1)
        if STAT(k, 1)==premium(i, 2)
            STAT(k, 2)=STAT(k, 2)+1;
            if premium(i, 4)==1
                STAT(k, 3)=STAT(k, 3)+1;
            end
        end
    end
end
STAT(k, 4)=STAT(k, 3)/STAT(k, 2);
end

赔付率计算代码:
clc;
statistical= xlsread('统计-赔付率.xlsx');
statistical(isnan(statistical)) = 0;
NUM_COUNTER=zeros(size(statistical, 1), 1);
STATISTIC_COUNTER=zeros(610, 13);
parameter=[3 5 7 9 11 13 15 16 19 21];
k=1;
for i=1:size(statistical, 1)
    if NUM_COUNTER(i)==0 %确认当前行是否被统计
        NUM_COUNTER(i)=1;
        STATISTIC_COUNTER(k, 1)=statistical(i, parameter(1));
        STATISTIC_COUNTER(k, 2)=statistical(i, parameter(2));
        STATISTIC_COUNTER(k, 3)=statistical(i, parameter(3));
        STATISTIC_COUNTER(k, 4)=statistical(i, parameter(4));
        STATISTIC_COUNTER(k, 5)=statistical(i, parameter(5));
        STATISTIC_COUNTER(k, 6)=statistical(i, parameter(6));
        STATISTIC_COUNTER(k, 7)=statistical(i, parameter(7));
        STATISTIC_COUNTER(k, 8)=statistical(i, parameter(8));
        STATISTIC_COUNTER(k, 9)=statistical(i, parameter(9));
        for j=1:size(statistical, 1) %统计数量

```

```

        if
STATISTIC_COUNTER(k,1)==statistical(j,parameter(1))&&STATISTIC_COUNTER(k,2)
==statistical(j,parameter(2))&&STATISTIC_COUNTER(k,3)==statistical(j,parameter(3))&&STATISTIC_COUNTER(k,4)==statistical(j,parameter(4))&&STATISTIC_COUNTER(k,5)==statistical(j,parameter(5))&&STATISTIC_COUNTER(k,6)==statistical(j,parameter(6))&&STATISTIC_COUNTER(k,7)==statistical(j,parameter(7))&&STATISTIC_COUNTER(k,8)==statistical(j,parameter(8))&&STATISTIC_COUNTER(k,9)==statistical(j,parameter(9))
            STATISTIC_COUNTER(k,11)=STATISTIC_COUNTER(k,11)+1;
            NUM_COUNTER(j)=1;
            if statistical(j,parameter(10))==1
                STATISTIC_COUNTER(k,12)=STATISTIC_COUNTER(k,12)+1;
            end
        end
    end
    STATISTIC_COUNTER(k,13)=STATISTIC_COUNTER(k,12)/STATISTIC_COUNTER(k,11)
;
    k=k+1;
end
end

x=zeros(size(STATISTIC_COUNTER,1),size(parameter,1));
for i=1:size(parameter,2)-1
    x(:,i)=STATISTIC_COUNTER(:,i);
end
y=STATISTIC_COUNTER(:,13);

```

计算年龄段与赔付率关系时代码:

```

clc;
statistical= xlsread('统计-赔付率.xlsx');
statistical(isnan(statistical)) = 0;
STATISTIC_COUNTER=zeros(8,4);
STATISTIC_COUNTER(:,1)=[0 1 2 3 4 5 6 7];
for k=1:size(STATISTIC_COUNTER,1)
    for i=1:size(statistical,1)
        if STATISTIC_COUNTER(k,1)==statistical(i,15)
            STATISTIC_COUNTER(k,2)=1+STATISTIC_COUNTER(k,2);
            if statistical(i,21)==1
                STATISTIC_COUNTER(k,3)=STATISTIC_COUNTER(k,3)+1;
            end
        end
    end
end
STATISTIC_COUNTER(k,4)=STATISTIC_COUNTER(k,3)/STATISTIC_COUNTER(k,2);
end

```

```
[p1,S1,mu1] = polyfit(STATISTIC_COUNTER(:,1),STATISTIC_COUNTER(:,4),5)
x=STATISTIC_COUNTER(:,1);
y=STATISTIC_COUNTER(:,4);
z=(-1.16e+09)+(-3.301e+11)*sin(-0.001504*x)+ (2.089e+09)*cos(2*-0.001504*x)
)+(2.641e+11)*sin(2*-0.001504*x)+(-9.284e+08)*cos(3*-0.001504*x)+ (-6.602e
+10)*sin(3*-0.001504*x);
plot(x,y,'X','color',[1 0 0])
hold on;
plot(x,z,'+','color',[0 0 1])
```