车辆保险分类费率厘定模型  
D1数据统计分析报告

|  |  |
| --- | --- |
| 版本号： | 0.2 |
| 修订日期： | 2013年08月10日 |
| 编写者： | 谭浩宇 |
| 公开范围： | 储月华、文军、谭浩宇、杨世栋 |
| 内容说明： | D1数据统计分析报告（基于HT-D1-v0.2.csv） |

# 目录

[目录 2](#_Toc363911193)

[1 简介 4](#_Toc363911194)

[2 分析结果 4](#_Toc363911195)

[2.1 基本数据观察与统计 4](#_Toc363911196)

[2.1.1 保单号（POLICYNO） 4](#_Toc363911197)

[2.1.2 险种代码（RISKCODE）与险别代码（KINDCODE） 4](#_Toc363911198)

[2.1.3 分公司（COMCODE） 5](#_Toc363911199)

[2.1.4 赔付次数（TIMES） 6](#_Toc363911200)

[2.1.5 赔付金额（SUMPAY）与案均赔付金额（AVGPAY） 8](#_Toc363911201)

[2.1.6 使用性质（USERNATURECODE） 16](#_Toc363911202)

[2.1.7 座位数（SEATCOUNT） 16](#_Toc363911203)

[2.1.8 排量（EXHAUSTSCALE） 16](#_Toc363911204)

[2.1.9 所属性质（PROPERTY） 17](#_Toc363911205)

[2.1.10 新车标志（NEWCARFLAG） 17](#_Toc363911206)

[2.1.11 行驶区域代码（RUNAREACODE） 17](#_Toc363911207)

[2.1.12 国产进口（COUNTRYNATURE） 17](#_Toc363911208)

[2.1.13 使用时间（USEYEARS） 17](#_Toc363911209)

[2.1.14 二手车标志（SECONDHANDCARFLAG） 18](#_Toc363911210)

[2.1.15 车辆类型（CARTYPE） 18](#_Toc363911211)

[2.1.16 车辆种类（CARKINDCODE） 18](#_Toc363911212)

[2.1.17 车辆品牌（BRANDNAME） 18](#_Toc363911213)

[2.1.18 实际价值（ACTUALVALUE）与新车购置价（PURCHASEPRICE） 18](#_Toc363911214)

[2.1.19 约定驾驶员（AGREEDDRIVERFLAG） 20](#_Toc363911215)

[2.1.20 同时投保车损三者（AB） 20](#_Toc363911216)

[2.1.21 业务来源（BUSINESSNATURE） 20](#_Toc363911217)

[2.1.22 续保标志（RENEWALFLAG） 21](#_Toc363911218)

[2.1.23 投保交强（COMPULSORY） 21](#_Toc363911219)

[2.1.24 承保月数（INSUREDMONTH） 21](#_Toc363911220)

[2.1.25 承保金额（AMOUNT） 21](#_Toc363911221)

[2.1.26 司机数量（DRIVERNUM） 21](#_Toc363911222)

[3 参考文献 22](#_Toc363911223)

# 简介

本报告以0.2版本的D1数据为基础，进行一系列基础统计分析，记录分析结果。主要目的是：1）进一步理解数据；2）观察数据质量；3）帮助完善D0至D1的数据预处理过程。

本报告使用的数据为储月华于2013年8月6日生成的D1数据，格式为CSV，文件名为HT-D1-v0.2.csv。本数据集有28个字段和137351条记录。字段名称与意义见表 1。

**注意：从2.1.3开始，所有分析将仅基于险种代码（RISKCODE）为1224且承保月数（INSUREDMONTH）精确等于12的保单，共112363条数据**。

表 ： 数据字段说明

|  |  |  |  |
| --- | --- | --- | --- |
| 字段名 | 字段意义 | 字段名 | 字段意义 |
| POLICYNO | 保单号 | **SECONDHANDCARFLAG** | 二手车标志 |
| RISKCODE | 险种代码 | **CARTYPE** | 车辆类型 |
| KINDCODE | 险别代码 | **CARKINDCODE** | 车辆种类 |
| COMCODE | 分公司 | **BRANDNAME** | 车辆品牌 |
| TIMES | 赔付次数 | **ACTUALVALUE** | 实际价值 |
| SUMPAY | 赔付金额 | **PURCHASEPRICE** | 新车购置价 |
| USENATURECODE | 使用性质 | **AGREEDRIVERFLAG** | 约定驾驶员 |
| SEATCOUNT | 座位数 | **AB** | 同时投保车损三者 |
| EXHAUSTSCALE | 排量 | **BUSINESSNATURE** | 业务来源 |
| PROPERTY | 所属性质 | **RENEWALFLAG** | 续保标志 |
| NEWCARFLAG | 新车标志 | **COMPULSORY** | 投保交强 |
| RUNAREACODE | 行驶区域代码 | **INSUREDMONTH** | 承保月数 |
| COUNTRYNATURE | 国产进口 | **AMOUNT** | 承保金额 |
| USEYEARS | 使用时间 | **DRIVERNUM** | 司机数量 |

# 分析结果

## 基本数据观察与统计

### 保单号（POLICYNO）

POLICYNO是保单的唯一编号，由19位数字组成。经检查，数据集中没有重复的POLICYNO。

### 险种代码（RISKCODE）与险别代码（KINDCODE）

RISKCODE是一个类别变量，在数据集中的分布与意义见表 2。

表 2：RISKCODE统计结果

|  |  |  |  |
| --- | --- | --- | --- |
| 变量值 | 数量 | 比例 | 意义 |
| 1224 | 118604 | 86.35% | 家用机动车险 |
| 1225 | 700 | 0.51% |  |
| 1226 | 9 | 0.01% |  |
| 1229 | 18038 | 13.13% | 家用机动车险(电销09版) |
| 共计 | **137351** | **100%** |  |

KINDCODE全部为A，表示所有保单均为机动车损失险。

由于RISKCODE为1224的保单占到了绝大多数，**以下分析将仅基于这部分数据中承保月数（INSUREDMONTH）精确等于12的保单，共112363条数据**。

### 分公司（COMCODE）

一共有28个分公司，各分公司保单数量分布不均匀，详情见图 1和图 2。排名前7位的分公司的保单数量占保单总数的70.32%（79012/112363）。

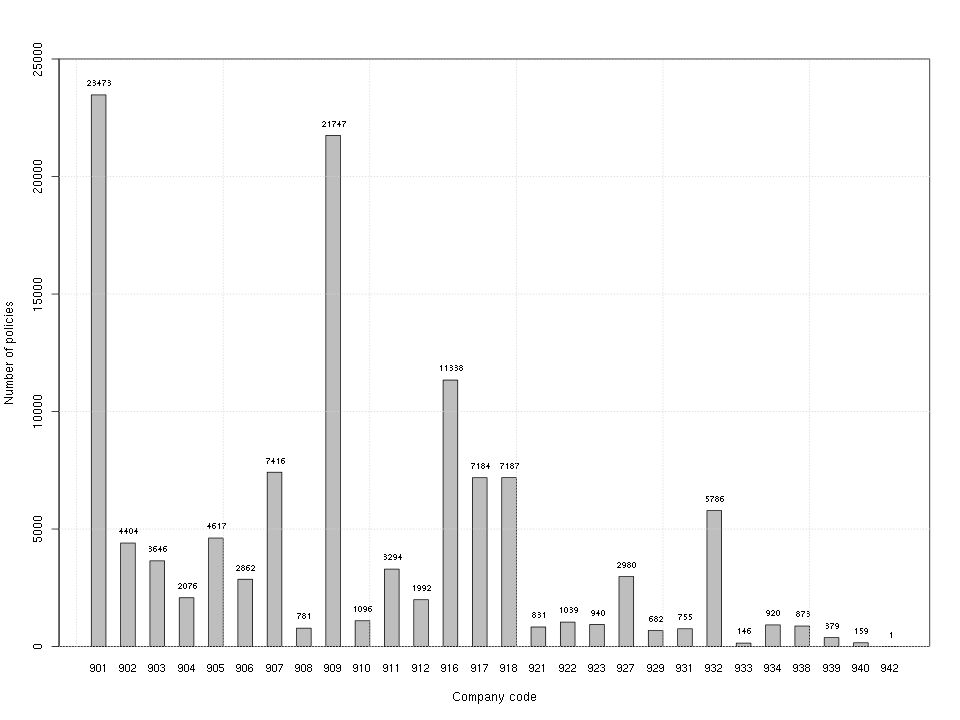


图 1：子公司保单数量分布。按照公司代码排序。

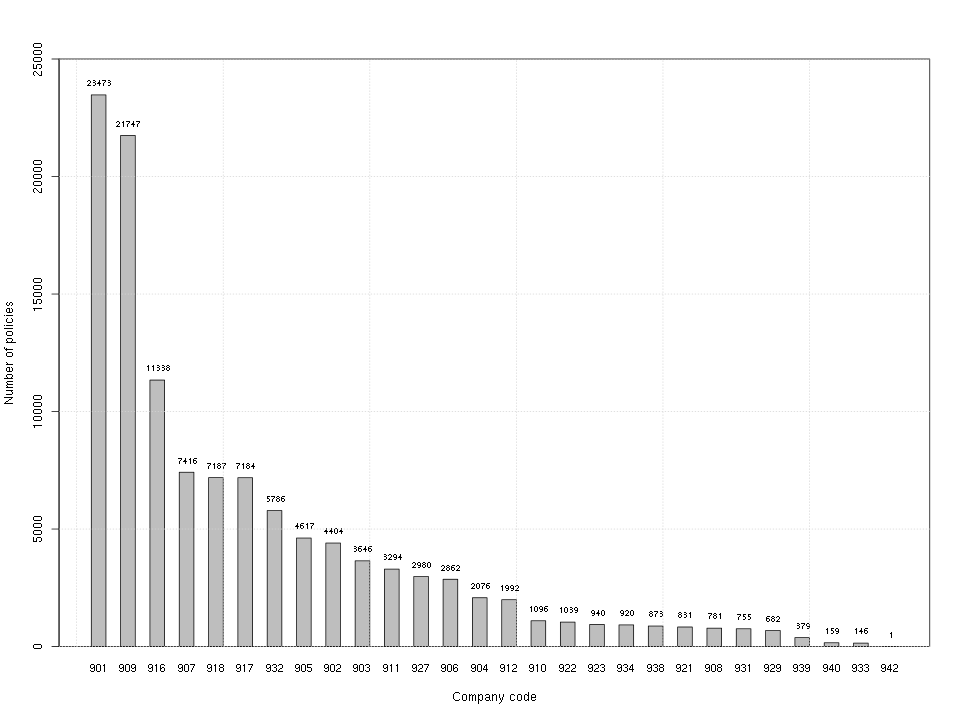


图 2：子公司保单数量分布。按照保单数量排序。

### 赔付次数（TIMES）

数据中，赔付次数分布在0至10之间，详细情况见图 3。其中出险次数为7、8、9、10的保单数量极少。理论上，赔付次数的概率分布应为泊松分布。图 4对赔付次数概率的经验分布与估计分布进行了对比。经验分布概率值的计算公式为:

由于赔付次数的平均值或方差分别为0.686和0.888，两者差距较大，可以判断出赔付次数与泊松分布拟合得并不是太好。当λ取两者中值0.787时，估计分布概率值的计算公式为：

可以看到，经验分布与估计分布有一定差距，尤其是在赔付次数为1时，估计值与实际值有较大差距。

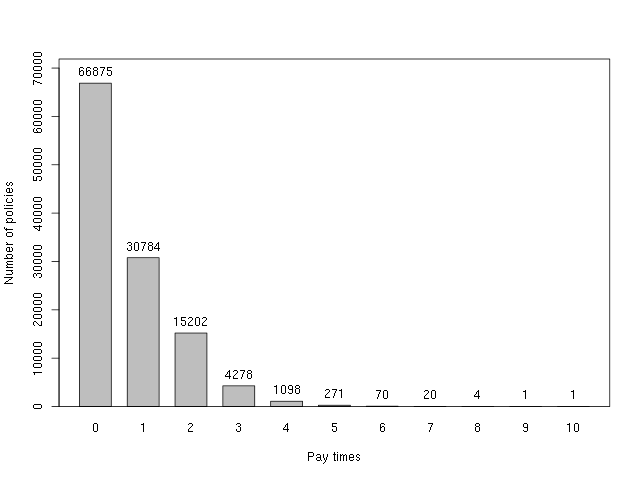


图 3：赔付次数分布

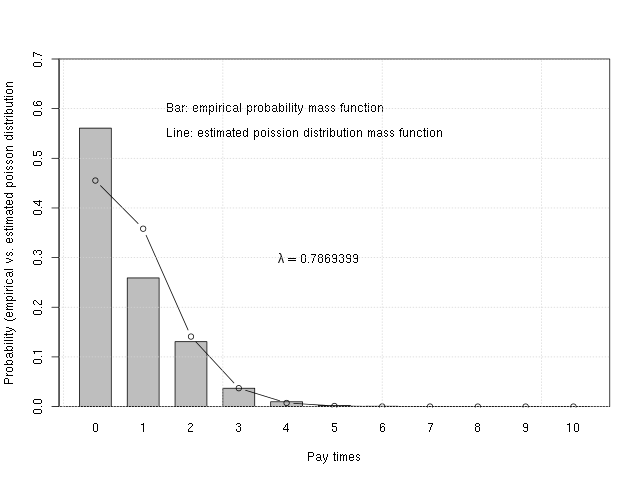


图 4：赔付次数经验概率分布与泊松概率分布对比

### 赔付金额（SUMPAY）与案均赔付金额（AVGPAY）

有赔付与无赔付的保单量对比见表 3。

表 3：有赔付与无赔付的保单量对比

|  |  |  |
| --- | --- | --- |
|  | 数量 | 比例 |
| SUMPAY = 0 | 63009 | 56.08% |
| SUMPAY > 0 | 49354 | 43.92% |
| 共计 | **118604** | **100%** |

经检查，赔付金额为0的保单，其赔付次数都为0，说明TIMES和SUMPAY这两个字段的一致性很好。

以我个人的理解，理论上，赔付金额应接近逆高斯分布或伽马分布，GLM对应这两种分布的连接函数要求响应变量为正数。也就是说，当对SUMPAY进行回归时，GLM无法处理SUMPAY为0的情况。因此，**下面的分析将仅针对赔付金额大于0的保单**。

讨论：文军找到的英文书[1]中也许有答案，需要进一步研究。

赔付金额和平均赔付的总体分布情况见图 5至图 8。为使观察更加细致，每个图分为8个子图，对应不同的赔付金额上限。例如，最后一个子图描述的是SUMPAY <= 5000时的数据分布情况。从图 7可以清楚地看到这份数据SUMPAY与AVGPAY的分布更接近逆高斯分布而不是伽马分布，这与 [1]中第2.8节的描述是吻合的。SUMPAY与AVGPAY的实际分布与估计分布的对比情况见图 9和图 10。

结论：针对这份数据，应用GLM预测赔付金额时应该采用逆高斯分布对应的连接函数。

另外还有一个有趣的现象。有些保单的赔付金额巨大，例如赔付金额前20位的保单，其赔付额都在200000以上，最大的一单达到了1288029.8元，但总体来看，小额赔付的总金额占全部赔付金额的大部分，具体情况见图 11。

结论：大额赔付看似金额巨大，但实际上没有小额赔付重要，再加上预测这些outliers本来也很困难，所以我们不必特别处理这些outliers，以免费力不讨好。

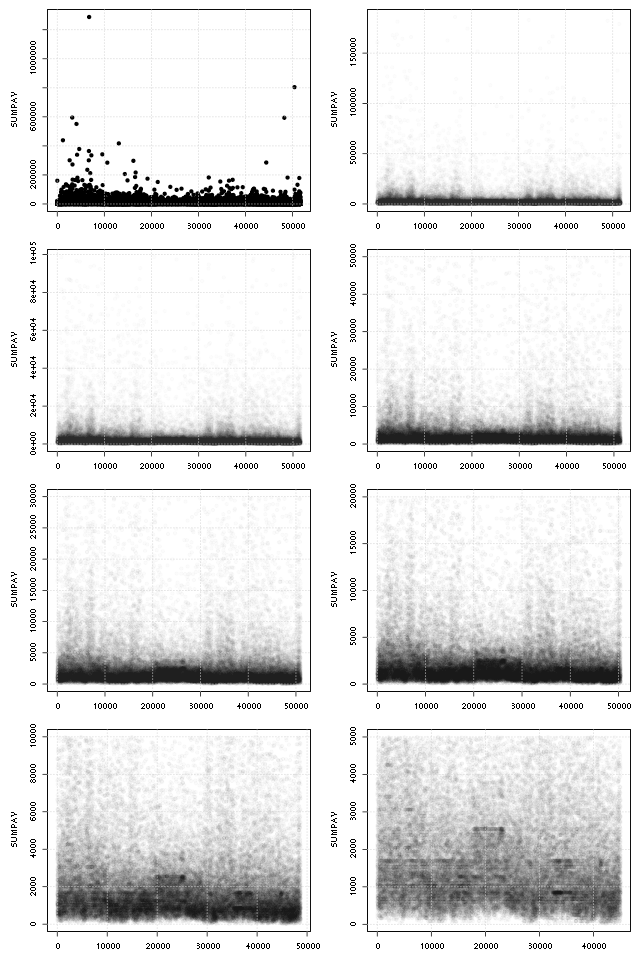


图 5：赔付金额总体分布情况（散点图）

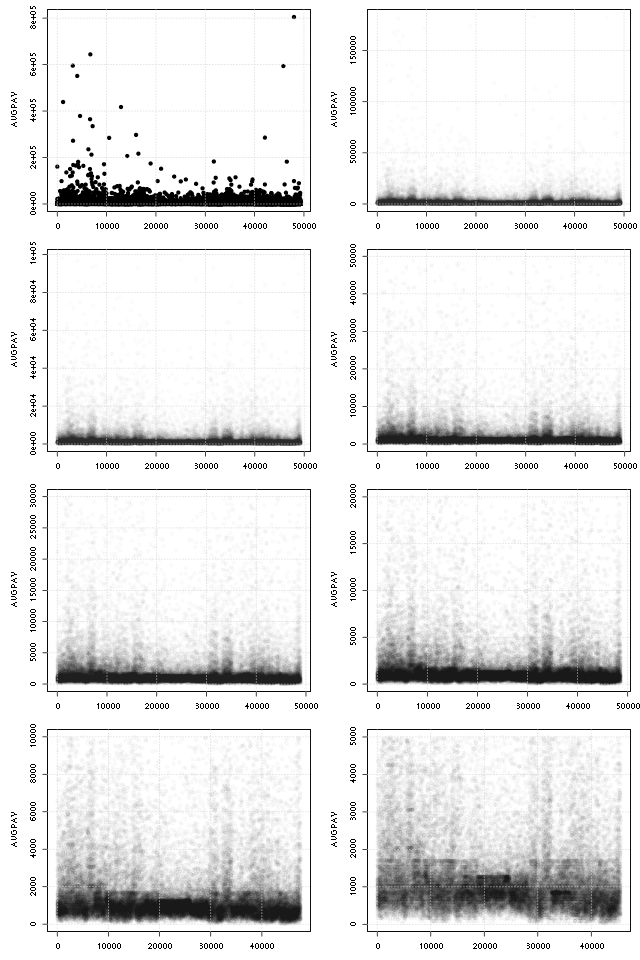


图 6：平均赔付金额总体分布情况（散点图）

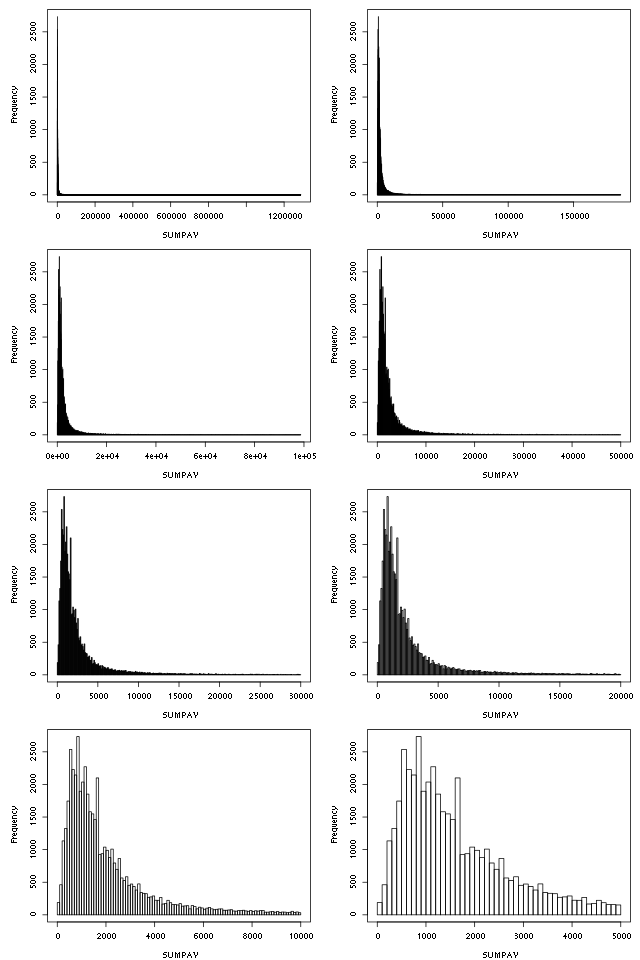


图 7：赔付金额总体分布（直方图）

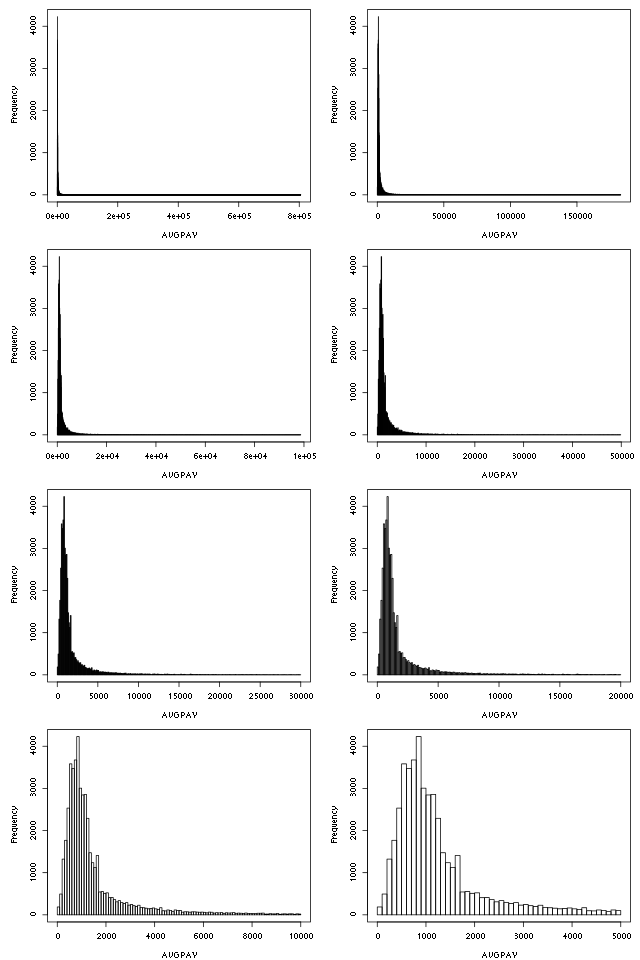


图 8：平均赔付金额总体分布（直方图）

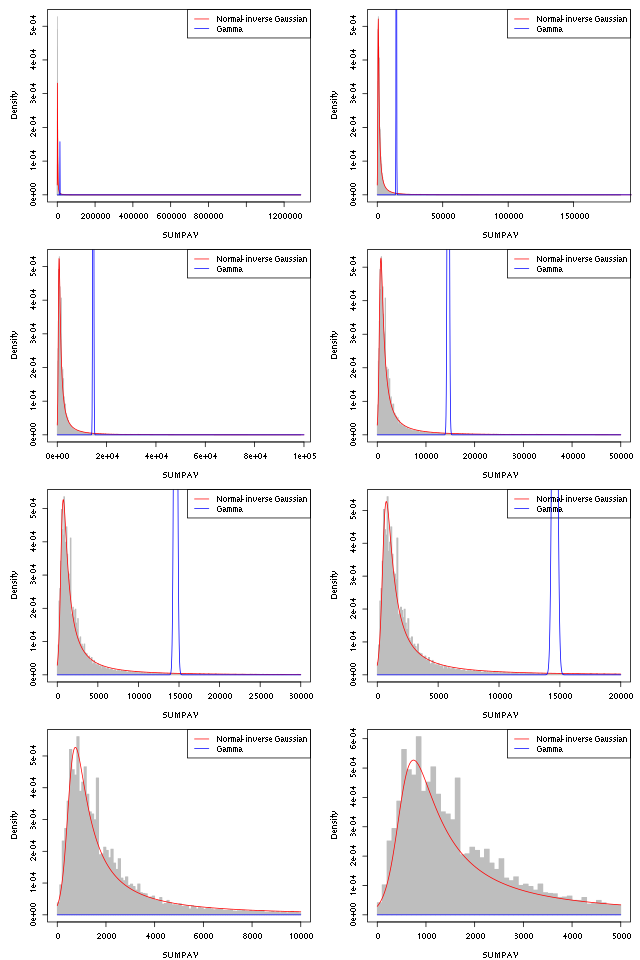


图 9：赔付金额实际分布与估计分布对比

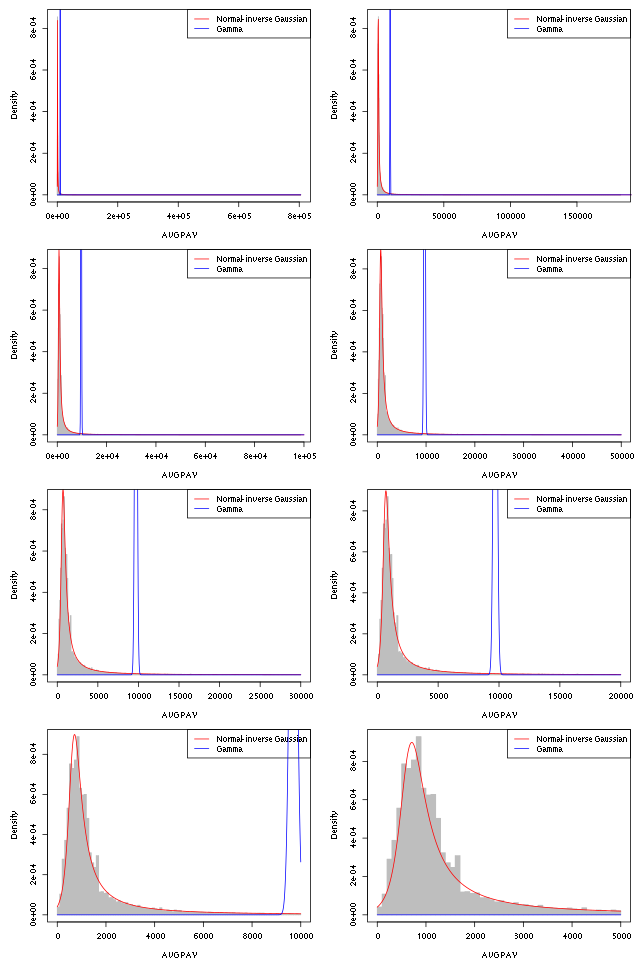


图 10：平均赔付金额实际分布与估计分布对比

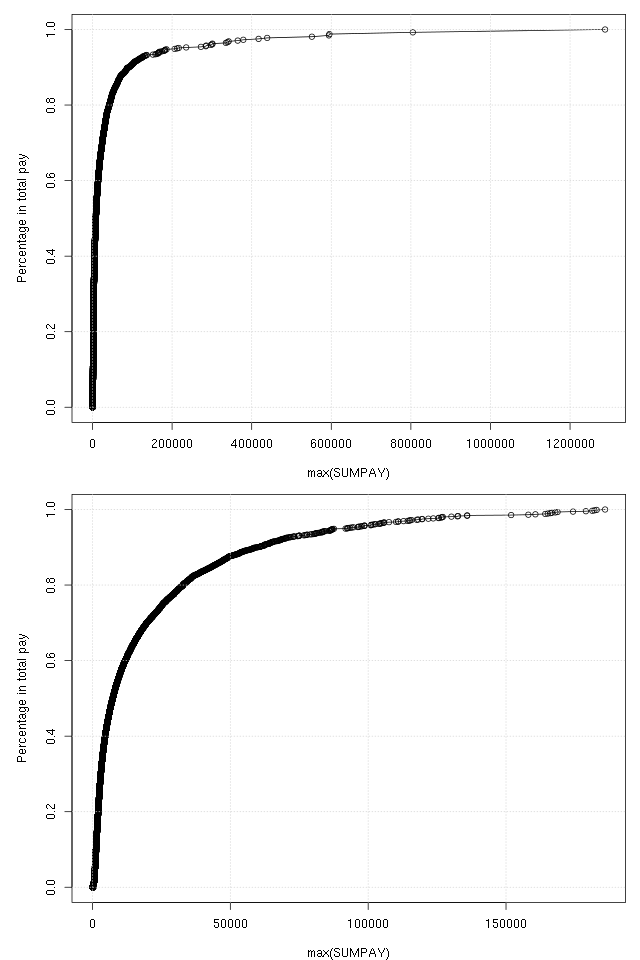


图 11：小额赔付总金额占全部赔付总金额比例。横坐标为赔付上限，纵坐标为所有低于赔付上限的赔付总金额占占全部赔付总金额比例。

### 使用性质（USERNATURECODE）

所有保单的使用性质均为8A，即家庭自用。

### 座位数（SEATCOUNT）

从表 4可以看到，绝大部分车辆的座位数是5、7、8，总比例约为97%，其中5座车的比例高达88%；有一小部分车辆座位数是4、6、9、11、2、10，总比例约为3%；其它车辆总比例不到0.1%。按照保监会分类，6座以下，6-10座，10座以上分类即可。

表 ：车辆座位数统计

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 2 | 3 | 4 | **5** | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 109 | 3 | 1516 | **98688** | 1144 | 7028 | 3174 | 414 | 61 | 140 | 15 |
| 13 | 14 | 15 | 17 | 18 | 19 | 32 | 35 | 37 | 47 | 57 |
| 2 | 30 | 16 | 7 | 3 | 8 | 1 | 1 | 1 | 1 | 1 |

### 排量（EXHAUSTSCALE）

车辆排量一般分布在0到8之间。有4辆车排量高达2021，判断为数据有误，应为2.021。其余车的排量分布见图 12。排量约为1.6、2.0和1.8的车辆数量最多。

讨论：在应用GLM时，这个字段应按照范围分段，具体分段方式需进一步研究。为了减少段数，是否可以采用非均匀分段方式？暂定为按照如下方式分段：(0, 1.0], (1, 1.5], (1.5, 1.6], (1.6, 1.8], (1.8, 2.0], (2.0, 3.0] , (3.0, INF]。

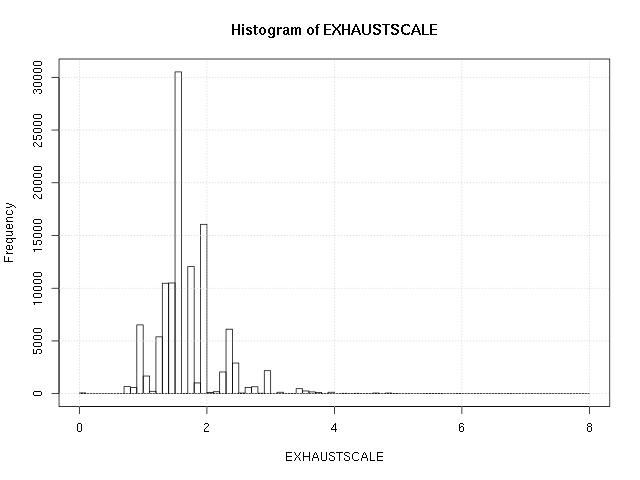


图 ：车辆排量分布

### 所属性质（PROPERTY）

车辆所属性质的分布见。其中， PROPERTY值04的保单比例达99.84%。04的意义是私人。

表 5：车辆所属性质分布

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 01 | 02 | 03 | 04 | 06 | 12 | 8A |
| 25 | 4 | 126 | 112188 | 2 | 17 | 1 |

### 新车标志（NEWCARFLAG）

该字段值为0或1。0代表旧车，数量为89724，比例为79.85%；1代表新车，数量为22639， 比例为20.15%。

### 行驶区域代码（RUNAREACODE）

这份数据中，该字段值仅有3和4两个值。3代表省内行驶，数量为13262，比例为11.80%；4代表国内行驶，数量为99101，比例为88.20%。

### 国产进口（COUNTRYNATURE）

该字段值的分布情况见表 6：国产进口分布。

表 ：国产进口分布

|  |  |  |
| --- | --- | --- |
| 字段值（意义） | 数量 | 比例 |
| A（国产车） | 6607 | 5．88% |
| B（进口车） | 31314 | 27.87% |
| C（合资车） | 74440 | 66.25% |
| NULL（其它） | 2 | 0% |
| 共计 | **112363** | **100%** |

### 使用时间（USEYEARS）

使用时间以年为单位，精确到1整年。从图 13可以看出，使用时间大体呈几何分布，即离散的指数分布，但使用年限为0的车辆数偏高。

使用时间为0的车辆有30874辆，但其中有8268辆没有标记为新车；此外，还有33辆车使用年限大于0，但被标记为新车，详情见表 7。

讨论：USEYEARS与NEWCARFLAG可能有些不一致，是否是因为USEYEARS本身应该为精确到月，小于9个月则认为是新车？我推测精确到月的使用时间是可以得到的，并且新车标志就是通过精确的使用时间计算得到的。

表 7：新车标志与使用年限对比

|  |  |  |  |
| --- | --- | --- | --- |
|  | USEYEARS == 0 | USEYEARS > 0 | 总数 |
| NEWCARFLAG == 1 | **22606** | **33** | 22639 |
| NEWCARFLAG == 0 | **8268** | **81456** | 89724 |
| 总数 | 30874 | 81489 | 112363 |

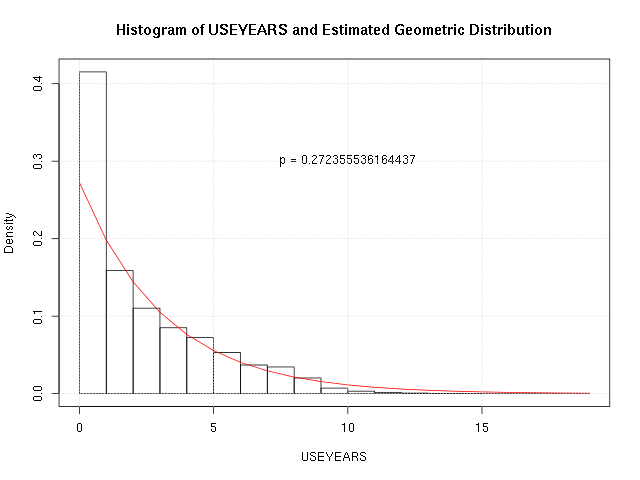


图 13：使用年限分布与估计几何分布

### 二手车标志（SECONDHANDCARFLAG）

该字段取值为0或1。1代表二手车，一共有1777辆，比例仅为1.58%，其中二手新车（SECONDHANDCARFLAG == 1 and NEWCARFLAG == 1）的数量为26。暂定建模时不使用该字段。

### 车辆类型（CARTYPE）

请储月华协助将该字段的取值处理为英文或数字。

### 车辆种类（CARKINDCODE）

所有保单的车辆种类都是A0，代表客车。

### 车辆品牌（BRANDNAME）

请储月华协助将该字段的取值处理为英文或数字。

### 实际价值（ACTUALVALUE）与新车购置价（PURCHASEPRICE）

通过观察图 14，可以发现实际价值（很可能）是通过一个简单公式，从新车购置价和使用时间计算得到的。计算方法是设定车辆折旧年限为11年，每年按照固定的比例降低实际价值，直到降低到新车购置价的20%为止。每年的折旧比例为0.8/11=0.727。因此，我们可以推测出下面这个公式：

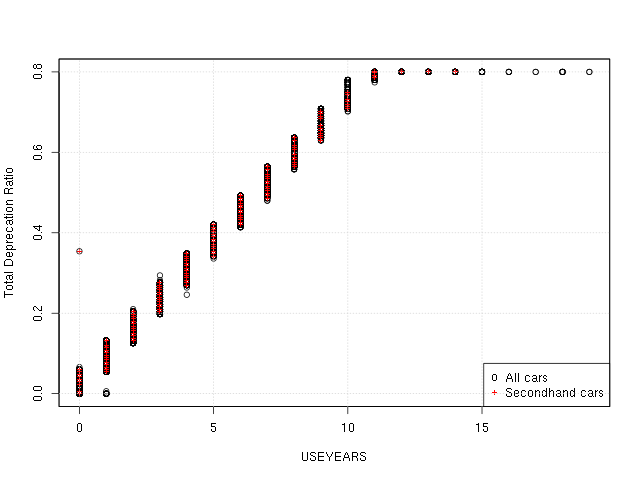


图 14：总折旧比例与使用时间的关系

另外可以看到，如果使用时间能够精确到月，则图 14中使用时间未超过11年的部分很可能会呈一条直线。

问题：是否能得到精确到月的使用时间？

若大部分的ACTUALVALUE是通过PRUCHASEPRICE和USEYEARS进行线性组合计算得来，则若后两个字段已经在模型中使用，加入ACTUALVALUE字段不会提升模型性能。

新车购置价的分布见图 15。

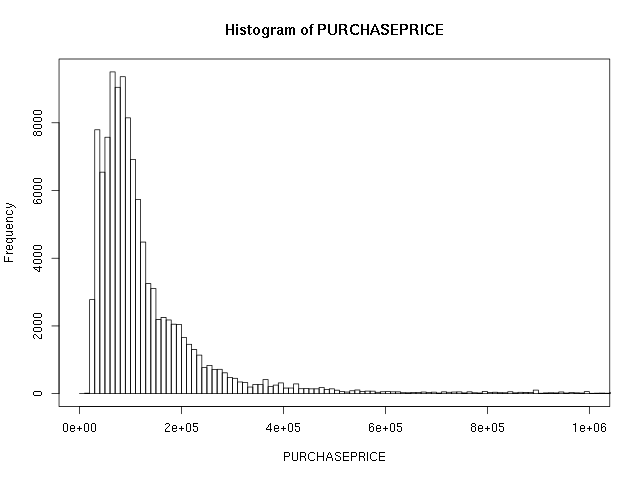


图 ：新车购置价分布

### 约定驾驶员（AGREEDDRIVERFLAG）

该字段取值为0或1。0代表无约定驾驶员，数量为102838， 比例为91.52%；1代表有约定驾驶员，数量为9525，比例为8.48%。

### 同时投保车损三者（AB）

该字段取值为A或AB。A代表仅车损险，数量为307，比例为0.27%；AB代表同时投保车损险和三者险，数量为112056， 比例为99.73%。

### 业务来源（BUSINESSNATURE）

该字段取值为0,1,2,3,4,6，具体分布见表 8。

表 ：业务来源分布

|  |  |  |
| --- | --- | --- |
| 字段值（意义） | 数量 | 比例 |
| 0直接业务 | 15333 | 13.65% |
| 1个人代理 | 4023 | 3.58% |
| 2专业代理 | 27122 | 24.14% |
| 3兼业代理 | 65221 | 58.04% |
| 4境内经纪业务 | 654 | 0.58% |
| 6电销/网销 | 10 | 0.01% |
| 共计 | **112363** | **100%** |

### 续保标志（RENEWALFLAG）

该字段取值为0,1,2，具体分布见表 9

商业续商业给1 交强续交强给2 否则给0

表 ：续保标志分布

|  |  |  |
| --- | --- | --- |
| 字段值（意义） | 数量 | 比例 |
| 0 新保险 | 70599 | 62.83% |
| 1 续保 | 41702 | 37.11% |
| 2 续保（可并入1） | 62 | 0.06% |
| 共计 | **112363** | **100%** |

### 投保交强（COMPULSORY）

该字段取值为0或1。0代表未在同一公司投保交强险，数量为6750，比例为6.01%；1代表在同一公司投保交强险，数量为105613，比例为93.99%。

### 承保月数（INSUREDMONTH）

本报告仅分析INSUREDMONTH精确等于12的保单。

### 承保金额（AMOUNT）

承保金额基本上与新车购置价保持一致，具体情况见表 10。

表 ：承保金额分布

|  |  |  |
| --- | --- | --- |
| X=承保金额/新车购置价 | 数量 | 比例 |
| X > 1.0 | 10 | 0.01% |
| X == 1.0 | 111898 | 99.58% |
| X < 1.0 | 402 | 0.36% |
| N/A | 53 | 0.05% |
| 共计 | **112363** | **100%** |

### 司机数量（DRIVERNUM）

该字段取值为0,1,2,3。0代表不限司机数量。从表 11可以看出DRIVERNUM与AGREEDDRIVEFLAG关系密切。

表 11：司机数量与约定司机对比

|  |  |  |  |
| --- | --- | --- | --- |
|  | DRIVERNUM == 0 | DRIVERNUM > 0 | 总数 |
| AGREEDDRIVEFLAG == 0 | **102838** | **0** | 102838 |
| AGREEDDRIVEFLAG == 1 | **86** | **9439** | 9525 |
| 总数 | 102924 | 9439 | 112363 |

问题：有86个保单约定了司机但没有指定司机数量，为什么？

# 参考文献

1. Generalized Linear Models for Insurance Data. Piet de Jong and Gillian Z. Heller.