湖南大學
HUNAN UNIVERSITY

统计仿真实验 2

实验报告

广义线性回归

车险数据分析

张笑竹
统计 1601
201618070114

1. 引言

1.1 背景介绍

随着汽车行业的繁荣，车险产品也得到了极大的发展。车险产品主要通过车因素、人因素和环境因素三个方面衡量被保险人的风险水平，从而确定保费。在车联网和大数据的不断推动下，一种新型车险逐渐走入人们的视线，它就是 UBI（Usage Based Insurance），即基于驾驶人行为的车险。UBI 模式车险是基于驾驶行为以及使用车辆相关数据相结合的个性化保险产品。

本案例通过对车险数据进行统计分析，建立出险因素 0-1 回归模型，挖掘影响出险的重要变量。在商业应用层面，本案例建立的出险因素模型对于制定个性化车险产品、识别不同风险的驾驶人具有一定的指导作用。在未来，结合驾驶行为数据，可制定基于驾驶行为的 UBI 车险产品。

1.2 数据集

本案例所使用的数据来自某保险公司提供的车险数据，包含 4233 条案例。变量包括是否出险（LossClass）、驾驶人年龄（Age）、驾驶人驾龄（exp）、驾驶人性别（Gender）、驾驶人婚姻状况（Marital）、汽车车龄（vAgeNew）、发动机引擎大小（EngSize）、是否进口（import）、所有者性质（owner）、固定车位（Garage）、防盗装置（AntiTFD）。

代码见附录 Markdown 文件。

2. 任务分析

2.1 任务一

在任务一中，读入数据并查看数据。11 个变量中，共有 7 个分类变量，5 个连续变量。在后续的分析中，“是否出险（LossClass）”将由连续变量转化为 0-1 变量，成为 Logistics 回归的响应因素。

```
##     EngSize          Age        Gender    Marital          exp
##  Min.   :1.000   Min.   :21.00   男:3775   未婚: 205   Min.   : 0.000
##  1st Qu.:1.600   1st Qu.:33.00   女: 458   已婚:4028   1st Qu.: 3.000
##  Median :1.800   Median :38.00                         Median : 5.000
##  Mean   :1.815   Mean   :38.18                         Mean   : 5.954
##  3rd Qu.:1.800   3rd Qu.:42.00                         3rd Qu.: 8.000
##  Max.   :3.000   Max.   :66.00                         Max.   :20.000
##   Owner      Garage          AntiTFD      import       LossClass
##  公司: 869   无: 687   无防盗装置:3285   国产:2970   Min.   :0.0000
##  私人:3027   有:3546   有防盗装置: 948   进口:1263   1st Qu.:0.0000
##  政府: 337                                          Median :0.0000
##                                                     Mean   :0.2847
##                                                     3rd Qu.:1.0000
##                                                     Max.   :1.0000
##  vAgeNew
##  旧车:1964
##  新车:2269
##
##
##
##
```

2.2 任务二

把变量“发动机引擎”进行离散化处理，处理办法参考目前国内轿车级别的分类标准，即 1.0-1.6 升为普通级车，1.6 以上为中高级车；绘制不同类型轿车的棘状图。

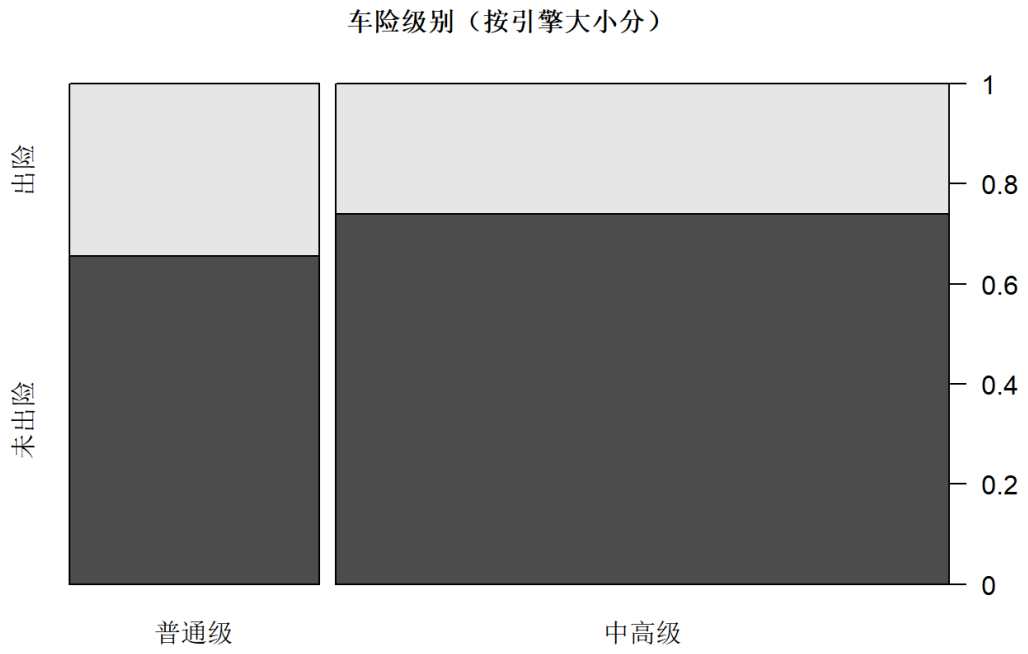


图 1　　不同类型轿车的棘状图

根据图 1，中高级车的出险率为 25%左右，而普通级车的出险率为 35%左右；因此中高级车的出险率较低，这也较为符合大众印象以及常识——中高级车的价格更贵，车况更好，车主使用习惯也往往更好。

当然，单纯地从汽车等级（引擎大小）的角度划分，出险率的差距并不十分明显；因此我们需要考察更多的因素进行判断。

2.3 任务三

在任务三中，我们将建立 Logistics 回归进行进一步分析。首先，我们构建全模型进行回归。根据拟合结果，在显著性水平为 0.05 时，“AntiTFD 有防盗装置”、“Age”、“Gender 女”和“Marital 已婚”并不显著。在这种情况下，我们需要通过加入模型复杂度的惩罚项来避免过拟合问题。

第一个调整模型为 AIC 模型，即以赤池信息量准则（AIC）筛选变量：

$$AIC = 2k + n\ln\left(\frac{RSS}{n}\right).$$

经过筛选，保留了 8 个变量，然而“Gender 女”仍然不显著。

第二个调整模型为 BIC 模型，即以贝叶斯信息准则（BIC）筛选变量。能够在不完全情报下，对部分未知的状态用主观概率估计，然后用贝叶斯公式对发生概率进行修正，最后再利用期望值和修正概率做出最优决策。公式为：

$$BIC = \ln(n)\,k - 2\ln(L)$$

其中，L为似然函数，$\ln(n)\,k$惩罚项在维数过大且训练样本数据相对较少的情况下，可以有效避免出现维度灾难现象。

经过筛选，BIC 模型保留了 5 个变量，且全部显著。

注意到，全模型、AIC 模型和 BIC 模型在变量选择和回归系数的估计上都存在差异，那么究竟应该保留哪一个模型？首先，绘制三个模型的 ROC 曲线：

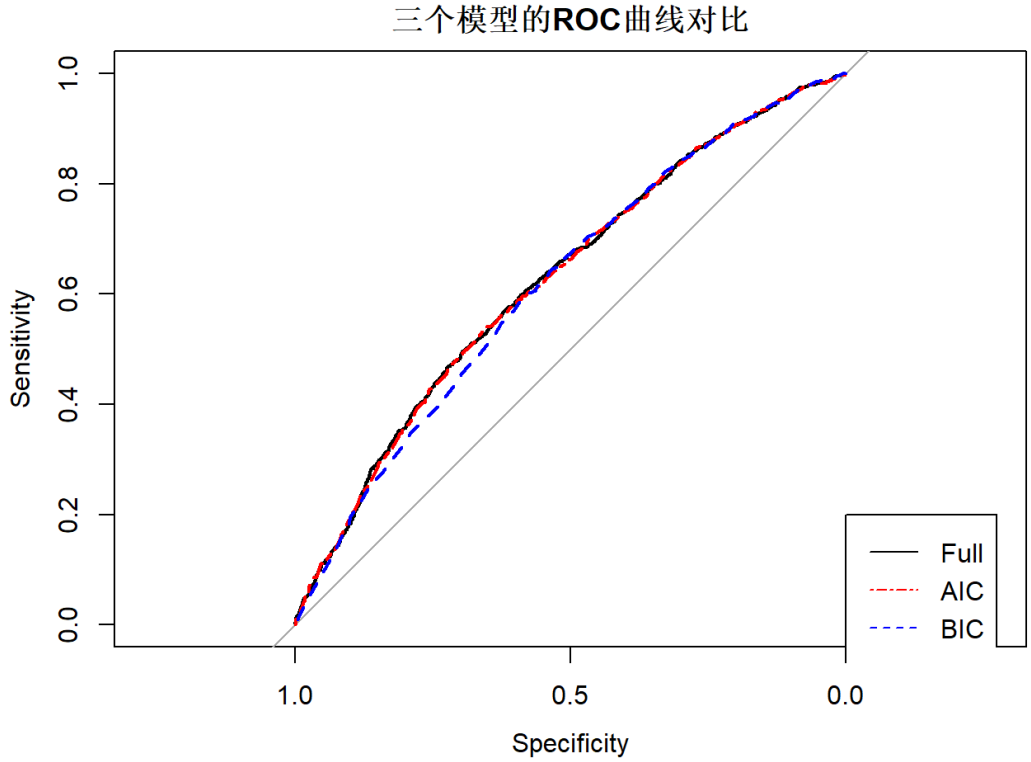


图 2　三个模型的 ROC 曲线对比

经过计算，三个模型的 AUC 值如下：

模型	全模型	AIC 模型	BIC 模型
AUC	0.6253	0.6241	0.6177

不难发现，全模型和 AIC 模型的 AUC 值相差无几，而 BIC 模型的 AUC 值则明显较小。虽然全模型的 AUC 值最大，但是考虑到减少过拟合的情况，AIC 模型应当是更好的选择。

2.4 任务四

在任务四中，我们选择了 AIC 模型。首先，在 ROC 曲线中遍历所有的阈值，并标出最佳阈值。

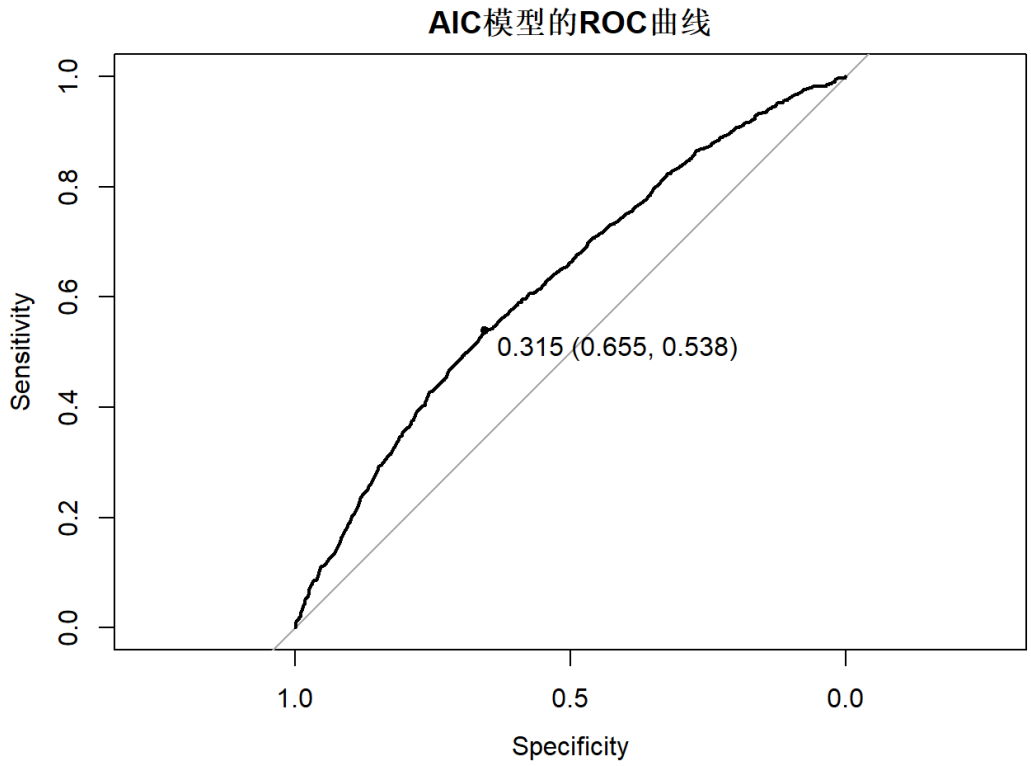


图 3　　AIC 模型的 ROC 曲线

当 Specificity 为 0.315，Sensitivity 为 0.538 时，达到了最佳的阈值 0.655。在此阈值下，对出险概率进行分配，并构建混淆矩阵（confusion matrix）：

```
##
##        0    1
##   0 2943 1164
##   1    6   10
```

正确率达到 71.62%，效果较好。

此外，我们需要对最终选定的 AIC 模型进行一些分析。通过观察模型的参数估计，不难发现，汽车新、有车库、进口车以及私家车更有可能造成交通事故，提高出险率；而对于高级车、政府用车以及较贵的车辆，出险率反而会降低。值得特别注意的是，从统计学意义上而言，女司机并不增加交通事故的风险。

2.5 任务五

最后，在任务五中，我们绘制出人群细分的柱形图，如图 4 所示。

事实上，上述的出险因素模型（AIC 模型）有一个十分有价值的应用领域：出险人群细分。大致做法是：首先按照 AIC 模型的预测出险概率进行从高到低排序，然后将排序后的驾驶人等分成 5 份，代表从高到底 5 种不同风险人群。将人群进行细分之后，可以计算这 5 种人群的实际出险概率。

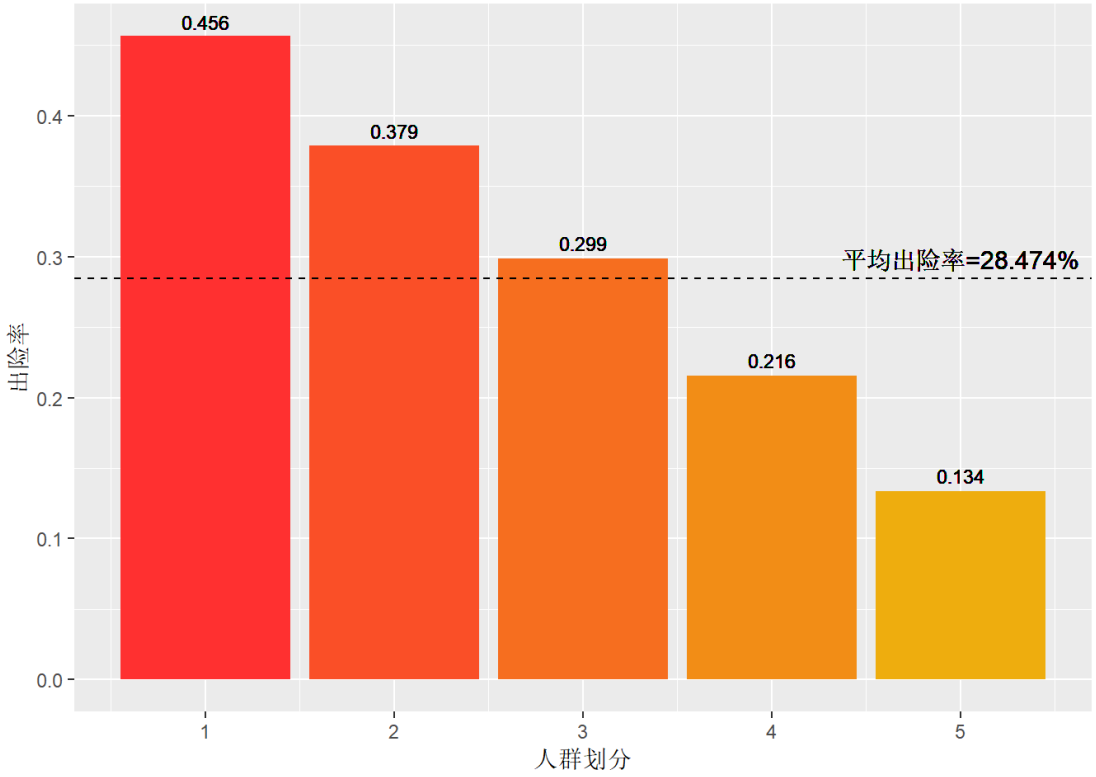


图 4　人群细分柱状图

通过细分人群，可以制定基于驾驶行为的 UBI 车险产品。如图 4 所示，平均出险率为 28.474%。这里，不妨进行如下规定：

$$\text{等级} = \begin{cases} \text{良好}, & \text{出险率} < 28.474\% \\ \text{不良}, & \text{出险率} \geq 28.474\% \end{cases}$$

对等级为“良好”的驾驶人，我们可以给予保费优惠；而对具有“不良”驾驶行为习惯的驾驶人，应适当提高保费。

3 结论

通过上述 5 个任务，我们不难发现，影响汽车保险出险率的因素有很多，其中，对于“汽车新”、“有车库”、“进口车”以及“私家车”的情况，保险公司应当提高警惕，因为符合此类特征的案例更有可能造成交通事故；而对于“高级车”、“政府用车”以及“较贵”的车辆，需要理赔的风险则相对较低。当然，从统计学意义上而言，女司机并不增加交通事故的风险，这一点并不仅仅有关于回归过程中过拟合，对于消除社会刻板印象而言，亦尤为重要。

当然，保险公司不能仅仅通过上述几个因素就主观判断该车的出险可能性，通过广义线性回归模型往往会得到更好的结果。通过使用我们在上面构建的 AIC 模型，并且选择 0.655 为阈值，对某一案例是否需要出险进行判断，最终能够实现至少 70%的准确性。

最后，我们可以根据预测的出险概率，从高到底细分出 5 种不同的风险人群。对于出险率低于平均水平 28.474%的人群，我们可以给予保费优惠；而对于出险

率高于平均水平的人群，应适当提高保费。

广义线性回归 - 车险数据分析

张笑竹 / 201618070114

2019年7月3日

随着汽车行业的繁荣，车险产品也得到了极大的发展。车险产品主要通过车因素、人因素和环境因素三个方面衡量被保险人的风险水平，从而确定保费。在车联网和大数据的不断推动下，一种新型车险逐渐走入人们的视线，它就是UBI（Usage Based Insurance），即基于驾驶人行为的车险。UBI模式车险是基于驾驶行为以及使用车辆相关数据相结合的个性化保险产品。本案例通过对车险数据进行统计分析，建立出险因素0-1回归模型，挖掘影响出险的重要变量。在商业应用层面，本案例建立的出险因素模型对于制定个性化车险产品、识别不同风险的驾驶人具有一定的指导作用。在未来，结合驾驶行为数据，可制定基于驾驶行为的UBI车险产品。

本案例所使用的数据来自某保险公司提供的车险数据，包括是否出险（LossClass）、驾驶人年龄（Age）、驾驶人驾龄（exp）、驾驶人性别（Gender）、驾驶人婚姻状况（Marital）、汽车车龄（vAgeNew）、发动机引擎大小（EngSize）、是否进口（import）、所有者性质（owner）、固定车位（Garage）、防盗装置（AntiTFD）。

1 任务一

读入数据并查看数据。

```
setwd("C:\\Users\\小竹子\\Desktop\\统计仿真实验\\4")
chexian <- read.csv("chexian.csv",header=T)
n <- nrow(chexian)
summary(chexian)
```

```
##     EngSize           Age        Gender    Marital          exp
##  Min.   :1.000   Min.   :21.00   男:3775   未婚: 205   Min.   : 0.000
##  1st Qu.:1.600   1st Qu.:33.00   女: 458   已婚:4028   1st Qu.: 3.000
##  Median :1.800   Median :38.00                         Median : 5.000
##  Mean   :1.815   Mean   :38.18                         Mean   : 5.954
##  3rd Qu.:1.800   3rd Qu.:42.00                         3rd Qu.: 8.000
##  Max.   :3.000   Max.   :66.00                         Max.   :20.000
##   Owner      Garage          AntiTFD      import       LossClass
##  公司: 869   无: 687   无防盗装置:3285   国产:2970   Min.   :0.0000
##  私人:3027   有:3546   有防盗装置: 948   进口:1263   1st Qu.:0.0000
##  政府: 337                                          Median :0.0000
##                                                     Mean   :0.2847
##                                                     3rd Qu.:1.0000
##                                                     Max.   :1.0000
##  vAgeNew
##  旧车:1964
##  新车:2269
##
##
##
##
```

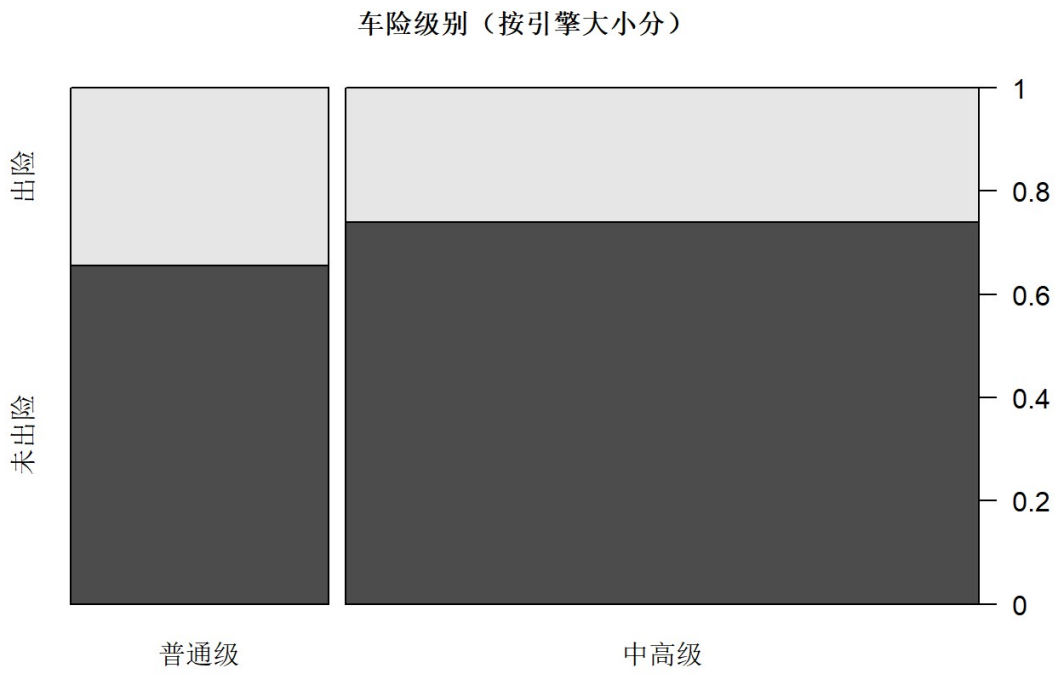
2 任务二

把变量’发动机引擎’进行离散化处理，处理办法参考目前国内轿车级别的分类标准，即1.0-1.6升为普通级车，1.6以上为中高级车；绘制不同类型轿车的棘状图。

```
EngSize <- c()
EngSize[which(chexian$EngSize > 1 & chexian$EngSize <= 1.6)] <- "普通级"
EngSize[which(chexian$EngSize > 1.6)] <- "中高级"
LossClass <- c()
LossClass[which(chexian$LossClass==1)] <- "出险"
LossClass[which(chexian$LossClass==0)] <- "未出险"
LossClass <- factor(LossClass,levels = c("未出险","出险"))
library(vcd)
```

```
## Loading required package: grid
```

```
EngSize <- factor(EngSize,levels = c("普通级","中高级"))
counts <- table(EngSize,LossClass)
spine(counts, main="车险级别（按引擎大小分）",xlab = "",ylab = "")
```



3 任务三

构建全模型、AIC模型和BIC模型，绘制ROC曲线；并根据曲线的AUC值选择模型。

（1）构造全模型

```
EngType<-EngSize
LossClass<-factor(chexian$LossClass)
chexian1<-data.frame(chexian[-10],EngType,LossClass)
summary(chexian1$LossClass)
```

```
##    0    1
## 3028 1205
```

```
glm1<-glm(LossClass ~ EngType + vAgeNew + AntiTFD + Garage + import + Owner + Age + exp + Gender + Mar
ital, family = binomial(link = "logit"), data = chexian1);glm1
```

```
##
## Call:  glm(formula = LossClass ~ EngType + vAgeNew + AntiTFD + Garage +
##     import + Owner + Age + exp + Gender + Marital, family = binomial(link = "logit"),
##     data = chexian1)
##
## Coefficients:
##       (Intercept)      EngType中高级         vAgeNew新车
##         -1.165428          -0.327460           0.371382
## AntiTFD有防盗装置            Garage有          import进口
##          0.088158           0.238682           0.153526
##          Owner私人          Owner政府                Age
##          0.315558          -0.374189          -0.003968
##               exp           Gender女        Marital已婚
##         -0.028718           0.198236           0.071448
##
## Degrees of Freedom: 4122 Total (i.e. Null);  4111 Residual
##   (110 observations deleted due to missingness)
## Null Deviance:       4926
## Residual Deviance: 4760  AIC: 4784
```

```
summary(glm1)
```

```
##
## Call:
## glm(formula = LossClass ~ EngType + vAgeNew + AntiTFD + Garage +
##     import + Owner + Age + exp + Gender + Marital, family = binomial(link = "logit"),
##     data = chexian1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2045  -0.8698  -0.7161   1.3371   2.2310
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -1.165428   0.255770  -4.557 5.20e-06 ***
## EngType中高级      -0.327460   0.080324  -4.077 4.57e-05 ***
## vAgeNew新车         0.371382   0.076470   4.857 1.19e-06 ***
## AntiTFD有防盗装置  0.088158   0.084558   1.043  0.29715
## Garage有            0.238682   0.102627   2.326  0.02003 *
## import进口          0.153526   0.080997   1.895  0.05803 .
## Owner私人           0.315558   0.100915   3.127  0.00177 **
## Owner政府          -0.374189   0.181324  -2.064  0.03905 *
## Age                -0.003968   0.005201  -0.763  0.44549
## exp                -0.028718   0.009138  -3.143  0.00167 **
## Gender女            0.198236   0.108418   1.828  0.06748 .
## Marital已婚         0.071448   0.170787   0.418  0.67570
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4926  on 4122  degrees of freedom
## Residual deviance: 4760  on 4111  degrees of freedom
##   (110 observations deleted due to missingness)
## AIC: 4784
##
## Number of Fisher Scoring iterations: 4
```

（2）构造AIC模型

```
glm2<-step(glm1)
```

```
## Start:  AIC=4783.97
## LossClass ~ EngType + vAgeNew + AntiTFD + Garage + import + Owner +
##     Age + exp + Gender + Marital
##
##           Df Deviance    AIC
## - Marital  1   4760.1 4782.1
## - Age      1   4760.6 4782.6
## - AntiTFD  1   4761.1 4783.1
## <none>         4760.0 4784.0
## - Gender   1   4763.3 4785.3
## - import   1   4763.5 4785.5
## - Garage   1   4765.5 4787.5
## - exp      1   4770.0 4792.0
## - EngType  1   4776.5 4798.5
## - Owner    2   4783.2 4803.2
## - vAgeNew  1   4783.8 4805.8
##
## Step:  AIC=4782.15
## LossClass ~ EngType + vAgeNew + AntiTFD + Garage + import + Owner +
##     Age + exp + Gender
##
##           Df Deviance    AIC
## - Age      1   4760.6 4780.6
## - AntiTFD  1   4761.2 4781.2
## <none>         4760.1 4782.1
## - Gender   1   4763.4 4783.4
## - import   1   4763.7 4783.7
## - Garage   1   4765.6 4785.6
## - exp      1   4770.1 4790.1
## - EngType  1   4776.6 4796.6
## - Owner    2   4783.4 4801.4
## - vAgeNew  1   4783.9 4803.9
##
## Step:  AIC=4780.6
## LossClass ~ EngType + vAgeNew + AntiTFD + Garage + import + Owner +
##     exp + Gender
##
##           Df Deviance    AIC
## - AntiTFD  1   4761.7 4779.7
## <none>         4760.6 4780.6
## - Gender   1   4764.1 4782.1
## - import   1   4764.2 4782.2
## - Garage   1   4766.1 4784.1
## - exp      1   4771.9 4789.9
## - EngType  1   4777.1 4795.1
## - Owner    2   4783.4 4799.4
## - vAgeNew  1   4784.3 4802.3
##
## Step:  AIC=4779.66
## LossClass ~ EngType + vAgeNew + Garage + import + Owner + exp +
##     Gender
##
##           Df Deviance    AIC
## <none>         4761.7 4779.7
## - Gender   1   4765.1 4781.1
## - import   1   4765.7 4781.7
## - Garage   1   4766.8 4782.8
## - exp      1   4773.0 4789.0
## - EngType  1   4777.6 4793.6
```

```
## - Owner    2   4784.7 4798.7
## - vAgeNew  1   4786.2 4802.2
```

```
summary(glm2)
```

```
##
## Call:
## glm(formula = LossClass ~ EngType + vAgeNew + Garage + import +
##     Owner + exp + Gender, family = binomial(link = "logit"),
##     data = chexian1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2028  -0.8703  -0.7147   1.3362   2.2244
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.221354   0.152651  -8.001 1.23e-15 ***
## EngType中高级 -0.320971   0.080044  -4.010 6.07e-05 ***
## vAgeNew新车    0.376283   0.076278   4.933 8.10e-07 ***
## Garage有       0.227793   0.102143   2.230 0.025739 *
## import进口     0.162766   0.080554   2.021 0.043324 *
## Owner私人      0.310716   0.100625   3.088 0.002016 **
## Owner政府     -0.369608   0.180677  -2.046 0.040787 *
## exp           -0.029634   0.008921  -3.322 0.000895 ***
## Gender女       0.203336   0.108070   1.882 0.059901 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4926.0  on 4122  degrees of freedom
## Residual deviance: 4761.7  on 4114  degrees of freedom
##   (110 observations deleted due to missingness)
## AIC: 4779.7
##
## Number of Fisher Scoring iterations: 4
```

（3）构造BIC模型

```
glm3<-step(glm1,k=log(n))
```

```
## Start:  AIC=4860.18
## LossClass ~ EngType + vAgeNew + AntiTFD + Garage + import + Owner +
##     Age + exp + Gender + Marital
##
##           Df Deviance    AIC
## - Marital  1   4760.1 4852.0
## - Age      1   4760.6 4852.4
## - AntiTFD  1   4761.1 4852.9
## - Gender   1   4763.3 4855.1
## - import   1   4763.5 4855.4
## - Garage   1   4765.5 4857.4
## <none>         4760.0 4860.2
## - exp      1   4770.0 4861.9
## - Owner    2   4783.2 4866.7
## - EngType  1   4776.5 4868.3
## - vAgeNew  1   4783.8 4875.6
##
## Step:  AIC=4852
## LossClass ~ EngType + vAgeNew + AntiTFD + Garage + import + Owner +
##     Age + exp + Gender
##
##           Df Deviance    AIC
## - Age      1   4760.6 4844.1
## - AntiTFD  1   4761.2 4844.7
## - Gender   1   4763.4 4846.9
## - import   1   4763.7 4847.2
## - Garage   1   4765.6 4849.2
## <none>         4760.1 4852.0
## - exp      1   4770.1 4853.6
## - Owner    2   4783.4 4858.5
## - EngType  1   4776.6 4860.1
## - vAgeNew  1   4783.9 4867.4
##
## Step:  AIC=4844.11
## LossClass ~ EngType + vAgeNew + AntiTFD + Garage + import + Owner +
##     exp + Gender
##
##           Df Deviance    AIC
## - AntiTFD  1   4761.7 4836.8
## - Gender   1   4764.1 4839.2
## - import   1   4764.2 4839.4
## - Garage   1   4766.1 4841.2
## <none>         4760.6 4844.1
## - exp      1   4771.9 4847.1
## - Owner    2   4783.4 4850.2
## - EngType  1   4777.1 4852.3
## - vAgeNew  1   4784.3 4859.5
##
## Step:  AIC=4836.81
## LossClass ~ EngType + vAgeNew + Garage + import + Owner + exp +
##     Gender
##
##           Df Deviance    AIC
## - Gender   1   4765.1 4832.0
## - import   1   4765.7 4832.5
## - Garage   1   4766.8 4833.6
## <none>         4761.7 4836.8
## - exp      1   4773.0 4839.8
## - Owner    2   4784.7 4843.1
```

```
## - EngType  1   4777.6 4844.4
## - vAgeNew  1   4786.2 4853.0
##
## Step:  AIC=4831.95
## LossClass ~ EngType + vAgeNew + Garage + import + Owner + exp
##
##           Df Deviance    AIC
## - import   1   4769.9 4828.3
## - Garage   1   4770.2 4828.7
## <none>         4765.1 4832.0
## - exp      1   4777.6 4836.1
## - Owner    2   4789.3 4839.4
## - EngType  1   4782.2 4840.6
## - vAgeNew  1   4790.6 4849.0
##
## Step:  AIC=4828.31
## LossClass ~ EngType + vAgeNew + Garage + Owner + exp
##
##           Df Deviance    AIC
## - Garage   1   4775.2 4825.3
## <none>         4769.9 4828.3
## - exp      1   4782.3 4832.4
## - EngType  1   4783.2 4833.3
## - Owner    2   4794.1 4835.8
## - vAgeNew  1   4799.9 4850.0
##
## Step:  AIC=4825.27
## LossClass ~ EngType + vAgeNew + Owner + exp
##
##           Df Deviance    AIC
## <none>         4775.2 4825.3
## - exp      1   4787.1 4828.9
## - EngType  1   4788.5 4830.2
## - Owner    2   4800.8 4834.2
## - vAgeNew  1   4807.6 4849.3
```

```
summary(glm3)
```

```
##
## Call:
## glm(formula = LossClass ~ EngType + vAgeNew + Owner + exp, family = binomial(link = "logit"),
##     data = chexian1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0667  -0.8794  -0.7238   1.3323   2.1458
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.01744    0.12968  -7.846  4.3e-15 ***
## EngType中高级 -0.27877    0.07599  -3.668 0.000244 ***
## vAgeNew新车    0.42402    0.07490   5.661  1.5e-08 ***
## Owner私人      0.32740    0.10029   3.264 0.001097 **
## Owner政府     -0.38509    0.18043  -2.134 0.032817 *
## exp           -0.03032    0.00887  -3.418 0.000630 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4926.0  on 4122  degrees of freedom
## Residual deviance: 4775.2  on 4117  degrees of freedom
##   (110 observations deleted due to missingness)
## AIC: 4787.2
##
## Number of Fisher Scoring iterations: 4
```

绘制ROC曲线，并根据曲线的AUC值选择模型。

```
library(gplots)
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

```
library(ROCR)
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
pred1<-predict(glm1,chexian1,type="response")
Roc1<-roc(chexian$LossClass,pred1)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

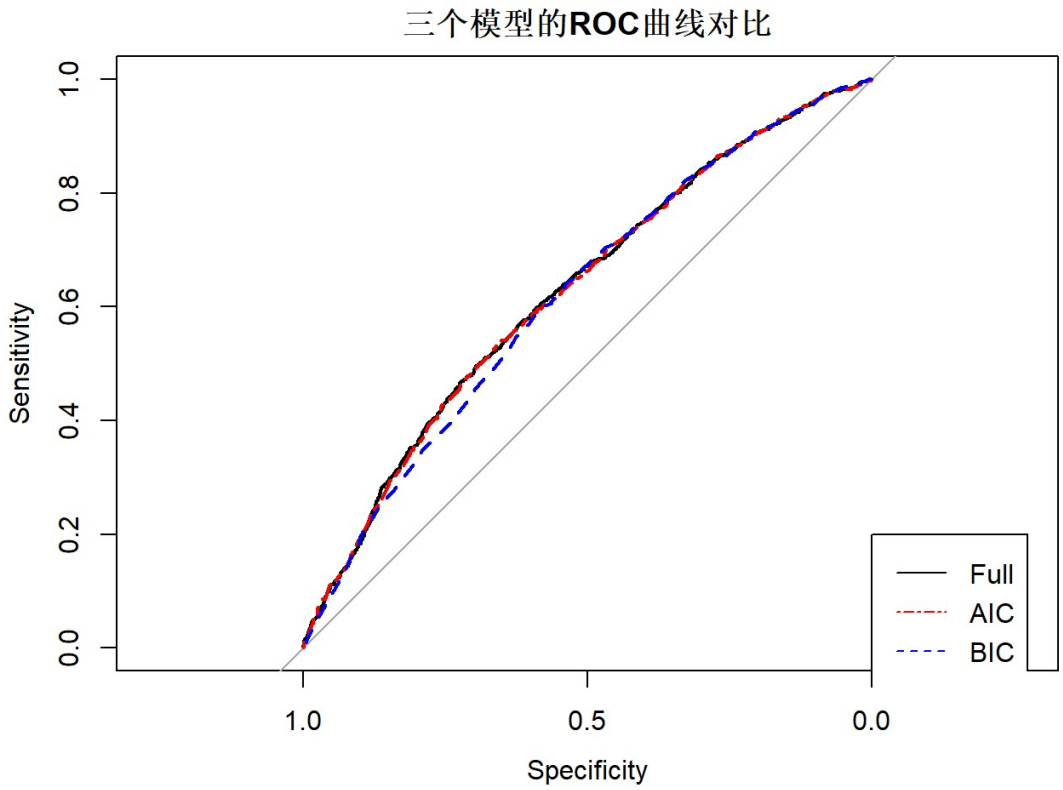
```
pred2<-predict(glm2,chexian1,type="response")
Roc2<-roc(chexian$LossClass,pred2)
```

```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```

```
pred3<-predict(glm3,chexian1,type="response")
Roc3<-roc(chexian$LossClass,pred3)
```

```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```

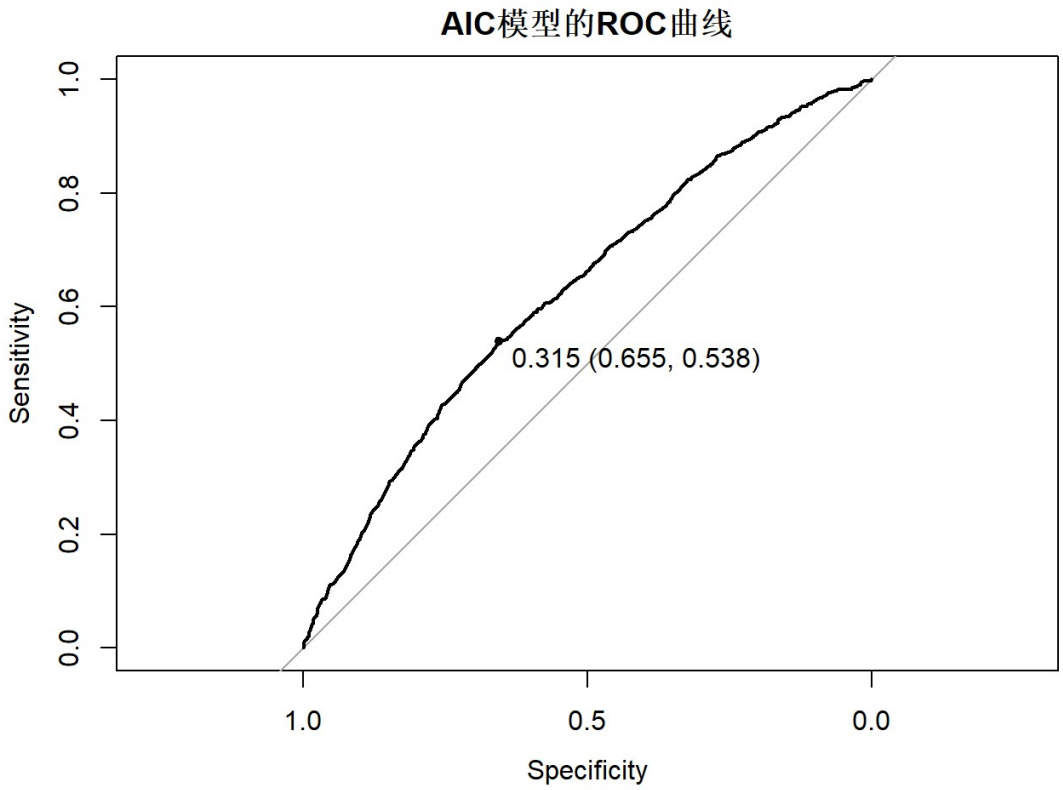
```
plot(Roc1,main="三个模型的ROC曲线对比")
plot(Roc2,add=TRUE,col="red",lty=6)
plot(Roc3,add=TRUE,col="blue",lty=2)
legend(0,0.2,inset = 0.5,c("Full","AIC","BIC"),lty=c(1,6,2),col=c("black","red","blue"))
```



4 任务四

选择AIC模型，在ROC曲线中显示最佳阈值，用最佳阈值来构建混淆矩阵。

```
library(gplots)
library(ROCR)
library(pROC)
plot(Roc2,main="AIC模型的ROC曲线",print.thres=TRUE)
```




```
auc<-auc(Roc2)
max(auc(Roc2))
```

```
## [1] 0.6263703
```

```
table(round(pred2),chexian1$LossClass)
```

```
##
##         0    1
##    0 2943 1164
##    1    6   10
```

```
pre<-c()
pre[which(pred2<=0.315)]<-0
pre[which(pred2>0.315)]<-1
pre<-factor(pre,levels = c(0,1))
LossClass11<-factor(chexian$LossClass,levels = c(0,1))
table(pre,LossClass11)
```

```
##     LossClass11
## pre    0    1
##    0 1900  538
##    1 1049  636
```

5 任务五

绘制人群细分的柱形图。

事实上，上述的出险因素模型的有一个十分有价值的应用领域：出险人群细分。大致做法是：首先按照AIC模型的预测出险概率进行从高到低排序，然后将排序后的驾驶人等分成5份，代表从高到底5种不同风险人群。将人群进行了细分之后，可以计算这5种人群的实际出险概率。

通过细分人群，可以制定基于驾驶行为的UBI车险产品，如对具有“良好”驾驶行为特征的驾驶人给予保费优惠，对具有“不良”驾驶行为习惯的驾驶人适当提高保费。

```
summary(pred2)
```

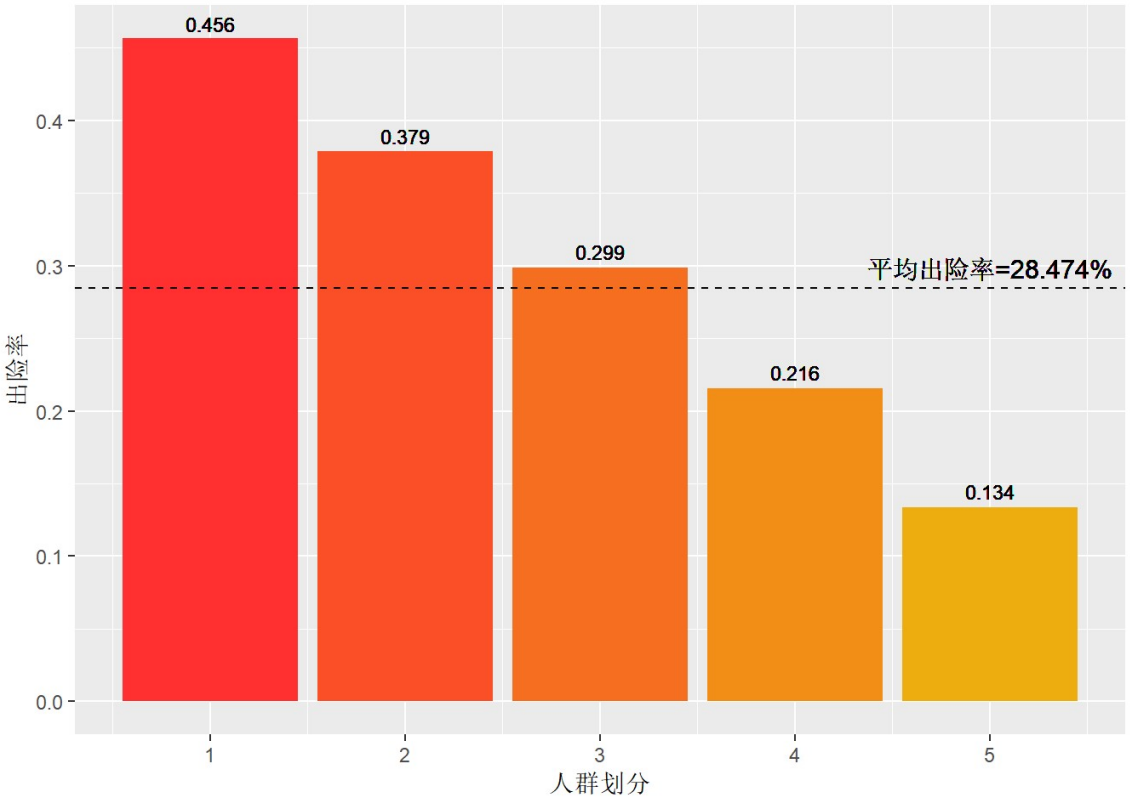
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
## 0.07554 0.22428 0.28899 0.28474 0.34904 0.51488     110
```

```
pred22=data.frame(pred2)
pred21=pred22[order(pred22$pred2),]
cutt=cut(-pred21,5,labels=F)
cut=data.frame(pred21,cutt)
library(ggplot2)
```

```
## Registered S3 methods overwritten by 'ggplot2':
##   method         from
##   [.quosures     rlang
##   c.quosures     rlang
##   print.quosures rlang
```

```
p=ggplot(cut,aes(x=cutt,y=pred21))
lable<-c(mean(pred21[which(cutt==1)]),mean(pred21[which(cutt==2)]),mean(pred21[which(cutt==3)]),mean(pred21[which(cutt==4)]),mean(pred21[which(cutt==5)]))
coo=colorRampPalette(c("firebrick1","darkgoldenrod2"))(5)
p<-p+stat_summary(geom = "bar",fun.y=mean,fill=coo[rank(1:5)])+geom_text(aes(x=1,y=0.456,label=0.456),vjust=-0.5,size=3)+geom_hline(aes( yintercept=0.28474),lty=2)+geom_text(aes(x=5,y=0.28474,label="平均出险率=28.474%"),vjust=-0.5,size=4)+geom_text(aes(x=2,y=0.379,label=0.379),vjust=-0.5,size=3)+geom_text(aes(x=3,y=0.299,label=0.299),vjust=-0.5,size=3)+geom_text(aes(x=4,y=0.216,label=0.216),vjust=-0.5,size=3)+geom_text(aes(x=5,y=0.134,label=0.134),vjust=-0.5,size=3)
p+labs(x="人群划分",y="出险率")
```

```
## Warning: Removed 110 rows containing non-finite values (stat_summary).
```


0.456
0.379
0.299
0.216
0.134
平均出险率=28.474%
出险率
0.0
0.1
0.2
0.3
0.4
1
2
3
4
5
人群划分