非寿险精算大作业二: 广义线性模型

在此次大作业中, 我们研究**年龄**和**车龄**对**车损险**索赔频率的影响.

第一步: 导入数据 "freq_data.csv". 对于字段的解释, 请参考课件 2. 此外, 该数据库包含以下新的字段:

- YEARS: 已赚风险单位数 (已赚车年数).
- Counts: 索赔次数 (出险次数, 报案次数).

第二步: 按表 1把年龄分为六组, 车龄分为四组. 在以后的分析中, 我们将使用这些分类变量代替连续型变量年龄和车龄.

Table 1: 年龄组和车龄组

年龄	年龄组	车龄	车龄组
[18, 24]	A	[0,1]	A
[25, 34]	B	[2, 4]	B
[35, 44]	C	[5, 8]	C
[45, 54]	D	[9, 14]	D
[55, 64]	E		
[65, 100]	F		

第三步: 假设索赔次数服从泊松分布, 且年龄和车龄没有交互作用. 把索赔次数作为因变量, 选取泊松分布的规范连接函数, 建立索赔频率的乘法模型.

问题:

1. (10 分) 参照课件, 写出模型及其基本假设.
2. (10 分) 计算参数的极大似然估计. 检验泊松分布的离散系数.
3. (10 分) 在显著水平 $\alpha = 0.05$ 下, 进行假设检验 H_0: 年龄组 E 和年龄组 F 的索赔频率没有差异.
4. (10 分) 在显著水平 $\alpha = 0.05$ 下, 进行假设检验 H_0: 车龄组 A 和车龄组 B 的索赔频率没有差异.
5. (10 分) 对于含有 100 个一年期保单的保单组, 假设投保人年龄均在 30 岁到 33 岁之间, 车龄均小于 1 年. (a) 预测该保单组的平均索赔次数, 及其预测均方误差. (b) 对于一份在该保单组的保单, 预测其索赔次数, 及其预测均方误差.

第四步: 按年龄组和车龄组对已赚风险单位数和索赔次数进行累积, 建立一个含有四个字段新数据库. 在新数据库中, 每一行对应一个年龄组和车龄组的组合, 四个字段分别为年龄组, 车龄组, 累积已赚风险单位数和累积索赔次数.

第五步: 假设索赔次数服从泊松分布, 且年龄和车龄没有交互作用. 把索赔次数作为因变量, 选取泊松分布的规范连接函数, 建立索赔频率的乘法模型.

问题:

6. (10 分) 证明: 建立在个体保单数据上的索赔次数模型 (第三步) 等价于建立在累积数据上的索赔次数模型 (第五步). 提示: 考虑似然函数.
7. (10 分) 计算参数的极大似然估计. 检验泊松分布的离散系数. 和第 2 问的答案进行对比.
8. (10 分) 如果考虑年龄和车龄的交互作用, 讨论模型参数的个数及其极大似然估计.

注意: 请用文字和数据回答以上问题, 不能直接粘贴 R 的输出结果. 请把相关代码作为附录.