第 7 讲：分类费率厘定 3 - 索赔频率模型

高光远

中国人民大学 统计学院

主要内容

- 假设索赔次数 N 服从泊松分布, 索赔频率为 λ, 车年数为 v. 我们想引入不同风险集合的**结构性差异**, 进而更准确地估计不同风险集合的索赔频率.
- 根据费率因子, 被保险人被划分在不同的风险集合. 假设有一个 d 维协变量空间 (费率因子空间) $\boldsymbol{x} = (x_1, \ldots, x_d)' \in \mathcal{X}$, 索赔频率回归方程 $\lambda(\cdot)$ 为一个映射 (mapping):

$$\lambda : \mathcal{X} \to \mathbb{R}_+, \qquad \boldsymbol{x} \mapsto \lambda = \lambda(\boldsymbol{x}).$$

- N 的分布为

$$N \sim \mathrm{Poi}(\lambda v)$$

- 定义平均索赔次数随机变量 $Y = N/v$.

N 的分布

可以把泊松分布转化为指数型分布族形式:

$$
\begin{aligned}
\Pr(N=k) &= \exp\left[-\lambda(\boldsymbol{x})v\right]\frac{(\lambda(\boldsymbol{x})v)^k}{k!} \\
&= \exp\left[\frac{k\log(\lambda(\boldsymbol{x})v)-\lambda(\boldsymbol{x})v}{1}-\log k!\right]
\end{aligned}
\tag{1}
$$

可知, $\theta=\log(\lambda(\boldsymbol{x})v), b(\theta)=\exp(\theta), c(k,\phi)=-\log k!, a(\phi)=1.$

Y 的分布

可以对平均索赔次数随机变量 $Y = N/v$ 建模, 其分布也为 EDF

$$\begin{aligned}\Pr(Y = k/v) &= \Pr(N = k) \\ &= \exp\left[-\lambda(\boldsymbol{x})v\right] \frac{(\lambda(\boldsymbol{x})v)^k}{k!} \\ &= \exp\left[\frac{\frac{k}{v}\log\lambda(\boldsymbol{x}) - \lambda(\boldsymbol{x})}{\frac{1}{v}} - \log k! + k\log v\right]\end{aligned} \tag{2}$$

可知 $\theta = \log\lambda(\boldsymbol{x}), b(\theta) = \exp(\theta), c(k,\phi) = -\log k! + k\log v$, $a(\phi) = 1/v$. 注意: Y 不服从泊松分布.

因为 $c(k,\phi)$ 对 β 的估计没有影响, 在求 β 的极大似然估计时, 可以假设 Y 服从期望为 $\lambda(\boldsymbol{x})$ 的泊松分布, 其权重为 v.

定义如下数学符号

$$
\begin{aligned}
\mathcal{D} &= \{(N_1, \boldsymbol{x}_1, v_1), \dots, (N_n, \boldsymbol{x}_n, v_n)\} \\
\beta &= (\beta_0, \dots, \beta_d)' \in \mathbb{R}^{d+1} \\
\log \lambda(\boldsymbol{x}) &= \beta_0 + \beta_1 x_1 + \cdots + \beta_d x_d = \langle \beta, \boldsymbol{x} \rangle \\
\boldsymbol{N} &= (N_1, \dots, N_n)' \\
X &= (x_{il})_{i=1:n, l=0:d} \in \mathbb{R}^{n \times (d+1)} \\
V &= \operatorname{diag}(v_1, \dots, v_n)
\end{aligned}
$$

极大似然估计 $\hat{\beta}$ 为下面方程的解

$$
X^T V \exp\{X\beta\} = X^T \boldsymbol{N} \tag{3}
$$

可通过 Newton-Raphson 算法, Fisher's scoring 方法, IRLS 方法计算上述方程的解 $\hat{\beta}$.

预测

- 索赔频率可**估计**为:

$$\hat{\lambda}(\boldsymbol{x}) = \exp\langle\hat{\beta}, \boldsymbol{x}\rangle$$

- $\hat{\lambda}(\boldsymbol{x})$ 的估计误差为

$$\mathrm{Var}(\hat{\lambda}(\boldsymbol{x})) \approx \hat{\lambda}(\boldsymbol{x})^2\mathrm{Var}(\hat{\eta}) = \hat{\lambda}(\boldsymbol{x})^2\boldsymbol{x}^T\mathrm{Var}(\hat{\beta})\boldsymbol{x} \tag{4}$$

- 平均索赔次数 $Y = N/v$ 可以通过 $\hat{\lambda}$ 进行**预测**:

$$\hat{Y} = \hat{\mathbb{E}}(Y) = \hat{\lambda}(\boldsymbol{x}).$$

- 假设偏差为零, 则预测均方误差为

$$\begin{aligned}\mathbb{E}\left[\left(Y_i - \hat{Y}_i\right)^2\right] &\approx \mathrm{Var}(\hat{Y}_i) + \mathrm{Var}(Y_i) \\ &\approx \hat{Y}_i^2\boldsymbol{x}_i^T\mathrm{Var}(\hat{\beta})\boldsymbol{x}_i + \frac{\hat{Y}_i}{v}\end{aligned} \tag{5}$$

- 可以看到, **过程方差和风险单位数成反比**.

残差

可以通过残差图评估**分布假设**和**连接函数假设**.

- Pearson 残差定义为

$$\epsilon_i^P = \frac{N_i - \hat{\lambda}(\boldsymbol{x}_i)v_i}{\sqrt{\hat{\lambda}(\boldsymbol{x}_i)v_i}}$$

- Deviance 残差定义为

$$\epsilon_i^D = \operatorname{sign}\left(N_i - \hat{\lambda}(\boldsymbol{x}_i)v_i\right)\sqrt{2N_i\left[\frac{\hat{\lambda}(\boldsymbol{x}_i)v_i}{N_i} - 1 - \log\left(\frac{\hat{\lambda}(\boldsymbol{x}_i)v_i}{N_i}\right)\right]}$$

如果 $N_i = 0$, 等式右边为 $\operatorname{sign}\left(N_i - \hat{\lambda}(\boldsymbol{x}_i)v_i\right)\sqrt{2\hat{\lambda}(\boldsymbol{x}_i)v_i}$.

偏差统计量

$$
\begin{aligned}
D(\beta_{full}, \hat{\beta}) &= D^*(\beta_{full}, \hat{\beta}) \\
&= \sum_{i=1}^{n} 2N_i \left[\frac{\hat{\lambda}(\boldsymbol{x}_i)v_i}{N_i} - 1 - \log\left(\frac{\hat{\lambda}(\boldsymbol{x}_i)v_i}{N_i} \right) \right] \\
&= \sum_{i=1}^{n} \left(\epsilon_i^D\right)^2
\end{aligned} \tag{6}
$$

如果 $N_i = 0$, 等式右边的第 i 项为 $2\hat{\lambda}(\boldsymbol{x}_i)v_i$.

离散系数

泊松模型中, 离散系数为常数 1. 这里需要检验因变量是否存在过离散 (over-dispersion)或者欠离散 (under-dispersion).

$$\begin{aligned} \hat{\phi}_P &= \frac{1}{n-d-1}\sum_{i=1}^{n}\left(\epsilon_i^P\right)^2 \\ \hat{\phi}_D &= \frac{1}{n-d-1}\sum_{i=1}^{n}\left(\epsilon_i^D\right)^2 \end{aligned} \tag{7}$$

$\hat{\phi}_P$ 和 $\hat{\phi}_D$ 应该接近于 1.

数据

使用第二周给出的保单数据库和理赔数据库, 这里只研究**交强险**的索赔次数.

考虑两个费率因子: **性别和年龄**. 其中, 性别为分类变量, 年龄为连续型变量. 拟解决如下几个问题:

1. 性别和年龄的交互作用 (interaction effect).
2. 对 N 建模和对 $Y = N/v$ 建模的等价性.
3. Deviance residuals VS Deviance statistics
4. 离散系数是否接近 1.
5. 假设检验: 性别对索赔频率没有显著的影响.
6. 对 30 岁男性驾驶员的平均索赔次数预测.

不考虑性别和年龄的交互作用

```
> ctp_poi<-glm(Counts~SEX+AGE,offset=log(YEARS),family=poisson(link="log"),data=data_ctp)
> summary(ctp_poi)

Call:
glm(formula = Counts ~ SEX + AGE, family = poisson(link = "log"),
data = data_ctp, offset = log(YEARS))

Deviance Residuals:
Min       1Q    Median       3Q      Max
-0.7343  -0.6896  -0.6576  -0.4204   4.6060

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.244963   0.121382 -10.257   <2e-16 ***
SEX2          0.047473   0.062438   0.760   0.4471
AGE         -0.005971   0.003077  -1.941   0.0523 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 4156.9  on 5840  degrees of freedom
Residual deviance: 4152.2  on 5838  degrees of freedom
AIC: 6316.5

Number of Fisher Scoring iterations: 6
```

注: 使用 offset 引入系数固定为 1 的协变量.

考虑性别和年龄的交互作用

```
> ctp_poi2<-glm(Counts~SEX*AGE,offset=log(YEARS),family=poisson(link="log"),data=data
      _ctp)
> summary(ctp_poi2)

Call:
glm(formula = Counts ~ SEX * AGE, family = poisson(link = "log"),
data = data_ctp, offset = log(YEARS))

Deviance Residuals:
Min        1Q    Median       3Q       Max
-0.7349  -0.6895  -0.6577  -0.4206   4.6060

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.246347   0.140254  -8.886   <2e-16 ***
SEX2         0.052478   0.261499   0.201   0.8409
AGE         -0.005934   0.003595  -1.651   0.0988 .
SEX2:AGE    -0.000137   0.006952  -0.020   0.9843
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 4156.9  on 5840  degrees of freedom
Residual deviance: 4152.2  on 5837  degrees of freedom
AIC: 6318.5

Number of Fisher Scoring iterations: 6
```

注：使用 * 考虑两个协变量的交互作用.

考虑性别和年龄的交互作用

- 不考虑交互作用. $\log\lambda$ 为截距不同, 斜率相同的两条线:

$$\log\lambda_i = \beta_0 + \beta_1 \mathbb{1}_{\mathrm{SEX}_i=\mathrm{Female}} + \beta_2 \mathrm{AGE}_i \tag{8}$$

男性: $\log\hat{\lambda} = -1.2450 - 0.0060 \times \mathrm{AGE}$
女性: $\log\hat{\lambda} = -1.1975 - 0.0060 \times \mathrm{AGE}$

- 考虑交互作用. $\log\lambda$ 为截距不同, 斜率不同的两条线:

$$\log\lambda_i = \beta_0 + \beta_1 \mathbb{1}_{\mathrm{SEX}_i=\mathrm{Female}} + \beta_2 \mathrm{AGE}_i + \beta_3 \mathbb{1}_{\mathrm{SEX}_i=\mathrm{Female}} \times \mathrm{AGE}_i \tag{9}$$

男性: $\log\hat{\lambda} = -1.2463 - 0.0059 \times \mathrm{AGE}$
女性: $\log\hat{\lambda} = -1.1938 - 0.0061 \times \mathrm{AGE}$

- 这里的交互作用指, 年龄对不同性别驾驶人的索赔频率的影响不同. 女驾驶员的索赔频率对年龄更加敏感 (但不统计显著).
- 由正态检验可知, 无法拒绝 $H_0 : \beta_3 = 0$.

对 N 建模和对 $Y=N/v$ 建模的等价性

```
> newY<-data_ctp$Counts/data_ctp$YEARS
> summary(glm(newY~SEX+AGE,weights =YEARS,family=poisson(link="log"),data=data_ctp))

Call:
glm(formula = newY ~ SEX + AGE, family = poisson(link = "log"),
    data = data_ctp, weights = YEARS)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7343  -0.6896  -0.6576  -0.4204   4.6060

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.244963   0.121382 -10.257   <2e-16 ***
SEX2         0.047473   0.062438   0.760   0.4471
AGE         -0.005971   0.003077  -1.941   0.0523 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 4156.9  on 5840  degrees of freedom
Residual deviance: 4152.2  on 5838  degrees of freedom
AIC: Inf

Number of Fisher Scoring iterations: 6
```

注：对平均索赔次数建模，需要引入 weights= 风险单位数.

Deviance residuals VS Deviance statistics

$$D(\beta_{full}, \hat{\beta}) = D^*(\beta_{full}, \hat{\beta}) = \sum_{i=1}^{n} \left(\epsilon_i^D\right)^2 \tag{10}$$

```
> deviance(ctp_poi)
[1] 4152.185
> sum(residuals.glm(ctp_poi,type="deviance")^2)
[1] 4152.185
```

注：两种残差, type=“deviance”, type=“pearson”.

离散系数是否接近 1

```
> deviance(ctp_poi)/(nrow(data_ctp)-length(ctp_poi$coefficients))
[1] 0.7112342
> sum(residuals.glm(ctp_poi,type="pearson")^2)/(nrow(data_ctp)-length(ctp_poi$
    coefficients))
[1] 1.161362
> summary(glm(Counts~SEX+AGE,offset=log(YEARS),family=quasipoisson(link="log"),data=
    data_ctp))$dispersion
[1] 1.161362
```

$$\hat{\phi}_D = 0.71, \hat{\phi}_P = 1.16.$$

注: 使用 `family=quasipoisson` 可以估计离散参数.

假设检验：性别对索赔频率没有显著的影响

```
> summary(ctp_poi)
Call:
glm(formula = Counts ~ SEX + AGE, family = poisson(link = "log"),
data = data_ctp, offset = log(YEARS))

Deviance Residuals:
Min        1Q   Median       3Q      Max
-0.7343  -0.6896  -0.6576  -0.4204   4.6060

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.244963   0.121382 -10.257   <2e-16 ***
SEX2         0.047473   0.062438   0.760   0.4471
AGE         -0.005971   0.003077  -1.941   0.0523 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
(Dispersion parameter for poisson family taken to be 1)
Null deviance: 4156.9  on 5840  degrees of freedom
Residual deviance: 4152.2  on 5838  degrees of freedom
AIC: 6316.5

Number of Fisher Scoring iterations: 6
> test_stat<-deviance(glm(Counts~AGE,offset=log(YEARS),family=poisson(link="log"),
     data=data_ctp))-deviance(ctp_poi)
> test_stat
[1] 0.574698
> qchisq(0.95, 1)
[1] 3.841459
> 1-pchisq(test_stat,df=1)
[1] 0.4483981
```

假设检验: 性别对索赔频率没有显著的影响

完整模型为:

$$\log \lambda_i = \beta_0 + \beta_1 \times \mathbb{1}_{\mathrm{SEX}_i=\mathrm{Female}} + \beta_2 \times \mathrm{AGE}_i$$

在 $H_0 : \beta_1 = 0$ 下的模型为

$$\log \lambda_i = \beta_0 + \beta_1 \times \mathrm{AGE}_i$$

正态检验: 从 `ctp_poi` 模型的 `summary` 表中可知, 检验统计量为 0.7600, p 值为 0.4471. 所以不能拒绝 H_0, 性别对索赔频率没有显著的影响.
似然比检验: 计算检验统计量为 0.5747. $\chi_1(95\%) = 3.8415$, p 值为 0.4483. 所以不能拒绝 H_0, 性别对索赔频率没有显著的影响.

对 30 岁男性驾驶员的平均索赔次数预测

```
> link_30<-predict(ctp_poi,newdata = data.frame(SEX="1",AGE=30,YEARS=1),type="link",
    se.fit = T) # linear prediction of log lambda
> response_30<-predict(ctp_poi,newdata = data.frame(SEX="1",AGE=30,YEARS=1),type="
    response",se.fit = T) # prediction of lambda
> exp(link_30$fit); response_30$fit; link_30$se.fit # exp (link) = response
        1
0.2407267
        1
0.2407267
[1] 0.04226038
> link_30$se.fit*response_30$fit; response_30$se.fit # the estimation  error
        1
0.0101732
        1
0.0101732
> sqrt(response_30$fit) # the process error
        1
0.4906391
> sqrt(response_30$se.fit^2+response_30$fit) # the MSEP for 1 risk exposure
        1
0.4907445
> sqrt(response_30$se.fit^2+response_30$fit/100) # the MSEP for 100 risk exposure
        1
0.0501075
```

对 30 岁男性驾驶员的平均索赔次数预测

$$\hat{Y} = 0.2407, \mathrm{Var}(\hat{\eta}) = 0.0423.$$

- 对于**一位 30 岁男性驾驶员**: 可知风险单位数为 1, 估计标准差为 0.0102, 过程标准差为 0.4906. 预测均方误差的平方根为 0.4907, 主要来源于**过程标准差**.
- 对于**含有一百位 30 岁男性驾驶员的风险集合**: 可知风险单位数为 100, 估计标准差为 0.0102, 过程标准差为 0.0491. 预测均方误差的平方根为 0.0501.

1. 证明式 (4) 和式 (5).