Generalised Linear Models

Guanxu Gleason WANG

University of Glasgow

版本：Past Paper of STAT5019 Generalised Linear Models, UofG

更新：April 7, 2025

目录

1 指数族分布与泊松回归

1.1 指数族分布 (Exponential Family of Distributions)

考虑一个随机变量 Y, 其概率密度函数 (p.d.f.) 或概率质量函数 (p.m.f.) 依赖于参数 θ. 该分布属于**指数族 (exponential family)**, 如果该分布可以写成

$$f(y;\theta) = \exp(a(y)b(\theta) + c(\theta) + d(y)), \tag{1}$$

其中 $b(\theta)$ 项被称为**自然参数 (natural parameter)**. 如果 $a(y) = y$, 那么这个分布也被称为**规范形式 (canonical form)**.

例 1.1 证明 Poisson(θ) 分布是指数族的成员之一, 且为规范形式.

证明. 由于 $Y \sim \text{Poisson}(\theta)$, 则它的 p.m.f. 为

$$f(y;\theta) = \frac{e^{-\theta}\theta^y}{y!},\ \theta > 0,\ y = 0, 1, 2, \cdots,$$

那么

$$\log f(y;\theta) = -\theta + y\log\theta - \log(y!).$$

识别参数:

$$a(y) = y, \quad b(\theta) = \log\theta, \quad c(\theta) = -\theta, \quad d(y) = -\log(y!).$$

因此 Poisson(θ) 分布是指数族的成员之一, 用例1.2中的结果计算数学期望得

$$\mathbb{E}[Y] = -\frac{c'(\theta)}{b'(\theta)} = -\frac{-1}{\frac{1}{\theta}} = \theta.$$

1.2 得分函数 (Score Function)

定义为对数似然函数关于参数 θ 的导数:

$$U(\theta) = \frac{\partial \log L(\theta)}{\partial \theta}$$

期望得分函数为零:

$$\mathbb{E}[U(\theta)] = 0$$

例 1.2 假设随机变量 Y 是指数族分布的规范形式, 定义单观察值的得分函数, 并利用其性质或其他方式证明

$$\mathbb{E}[Y] = -\frac{c'(\theta)}{b'(\theta)}.$$

解. 由于随机变量 Y 是指数族分布的规范形式, 因此可知它的 p.d.f. 或 p.m.f. 为

$$f(y;\theta) = \exp(yb(\theta) + c(\theta) + d(y)),$$

其对数似然函数为

$$\log L(\theta) = \sum [y_i b(\theta) + c(\theta) + d(y_i)],$$

计算得分函数为

$$U(\theta) = \frac{\partial \log L(\theta)}{\partial \theta} = \sum \left[y_i b'(\theta) + c'(\theta) \right].$$

取期望并令其等于 0, 即

$$\mathbb{E}\left[U(\theta)\right] = \mathbb{E}\left[Y\right] b'(\theta) + c'(\theta) = 0 \quad \Longrightarrow \quad \mathbb{E}\left[Y\right] = -\frac{c'(\theta)}{b'(\theta)}$$

1.3 构建合适的 GLM

广义线性模型 (Generalized Linear Model, GLM) 是线性回归的推广, 用于建模非正态分布的响应变量. 相比于传统的线性回归, GLM 允许响应变量服从指数族分布, 并通过链接函数 (Link Function) 建立预测变量和响应变量之间的关系.

表 1: 常见数据类型与分布选择及链接函数

响应变量类型	适用分布	误差结构	常用链接函数
连续变量	$\mathcal{N}(\mu, \sigma^2)$	方差恒定	恒等函数 $g(\mu) = \mu$
计数数据	$\text{Poisson}(\lambda)$	均值等于方差	对数函数 $g(\lambda) = \log(\lambda)$
二分类	$\text{Binomial}(n, p)$	方差为 $p(1-p)$	逻辑函数 $g(p) = \log\left(\frac{1-p}{p}\right)$
比例数据	$\text{Beta}(\alpha, \beta)$	依赖于参数	Logit 或 Probit

在 GLM 中，响应变量的分布可能不服从正态分布，因此需要链接函数将非线性关系变换为线性

关系:

$$g(\mathbb{E}[Y]) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots .$$

以 Poisson 分布为例: 由于泊松分布的性质 λ 必须是正值 (因为计数数据不能是负数), 而线性回归的右侧是一个无约束的线性组合, 可能是负值. 因此, 我们不能直接用线性回归建模, 即

$$\mathbb{E}[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots .$$

因为右侧可能导致 $\mathbb{E}[Y] < 0$ 的情况, 不符合泊松分布的定义.

此外, 由于模型的一些相关参数在不同种类下不同, 因此我们不能直接进行建模, 而是要标准化. 在泊松回归中, 我们通过偏置 (offset) 项来调整:

$$g(\mathbb{E}[Y]) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \log(\text{offset factor}).$$

下面以泊松回归为例, 引入两个例题: 例1.3 (食物摄入模式的分析) 和例1.4 (医院感染率的标准化分析).

例 1.3 Tomatoes, broccoli and carrots are three foods thought to contain cancer-fighting components. In a nutritional study, researchers attempt to investigate dietary patterns in a population, to understand the patterns of combined consumption of these components that occur. Subjects in a random sample of the population are surveyed and, based on their dietary histories, classified as high or low consumers of each of the three different foods. The results are described in a $2 \times 2 \times 2$ contingency table.

番茄、西兰花和胡萝卜被认为是含有抗癌成分的三种食物。在一项营养研究中，研究人员试图调查人群中的饮食习惯，以了解这些成分联合摄入的模式。对人群随机样本中的受试者进行调查，并根

据他们的饮食习惯，将他们分类为三种不同食物的高消费者或低消费者。结果描述在一个 $2 \times 2 \times 2$ 列联表中。

解. 我们研究的是不同人群摄入不同食品组合的个体数量, 即计数数据. 计数数据通常符合泊松分布:

$$\Pr(Y = y) = \frac{\lambda^y e^{-\lambda}}{y!}, \quad y = 0, 1, 2, \ldots$$

其中 λ 是该类别的期望计数, 即 $\mathbb{E}[Y] = \lambda$.

在泊松回归中，我们使用对数连接函数:

$$\log(\mathbb{E}[Y]) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

这个方程表示, 每个食品（番茄、花椰菜、胡萝卜）摄入情况对个体数量的影响是指数级的, 而不是简单的线性关系.

因此, 建立模型:

$$\log(\mathbb{E}[Y]) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

其中

- Y: 属于某一特定食物组合人群的计数
- X_i: $i = 1, 2, 3$ 分别是否高摄入番茄、花椰菜、胡萝卜 ($1 =$ 高, $0 =$ 低)
- β_0: 基准组（所有食物低摄入）的期望对数值
- $\beta_1, \beta_2, \beta_3$: 代表摄入番茄、花椰菜、胡萝卜对总人数的影响

例 1.4 Serious infections acquired during hospitalisation are recorded and counted over a five-year period for various hospitals in the US, classified as teaching or non-teaching, public or private, and religious vs secularly-run hospitals. Only some hospitals were in the study for the full five-year period. Also, the lengths of hospital stays vary considerably, and the hospitals are of different sizes with different numbers of occupied beds each day. For these reasons, it is decided to divide the count in each hospital by the average number of persons/day in the hospital, multiplied by the number of days that the hospital was in the study, before comparisons are made between different types of hospitals.

在美国，各种医院（分为教学医院和非教学医院、公立医院和私立医院、宗教医院与非宗教医院）在住院期间获得的严重感染被记录并计算，时间跨度为五年。只有部分医院参与了整个五年的研究。此外，住院时间长短差异很大，医院的规模不同，每天占用床位数量也不同。因此，决定在比较不同类型医院之前，将每家医院的感染计数除以医院平均每日人数，再乘以医院参与研究的天数。

解. 我们研究的是医院感染人数的计数数据, 符合泊松分布:

$$\Pr(Y = y) = \frac{\lambda^y e^{-\lambda}}{y!}, \quad y = 0, 1, 2, \ldots$$

计数数据的方差通常与均值相等（$\mathrm{Var}[Y] = \mathbb{E}[Y]$），符合泊松假设.

$$\log(\mathbb{E}[Y]) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

这样可以确保预测值始终为正, 并且可以解释为感染病例数相对于不同医院类别的变化.

由于医院规模不同, 住院天数不同, 我们不能直接比较感染人数, 而是要标准化:

$$\text{标准化感染数} = \frac{\text{感染病例数}}{\text{平均每日病床数} \times \text{医院参与天数}}$$

在泊松回归中, 我们通过偏置 (offset) 项 来调整:

$$\log(\mathbb{E}[Y]) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \log(\text{病床数} \times \text{天数})$$

这样可以确保模型的预测值是标准化的感染率, 而不是原始感染人数.

因此, 建立模型:

$$\log(\mathbb{E}[Y]) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \log(\text{病床数} \times \text{天数})$$

其中

- Y: 医院在某段时间内的感染病例数
- X_i: $i = 1, 2, 3$ 分别表示教学/非教学、公立/私立、宗教/世俗医院 ($1 =$ 是, $0 =$ 非)
- β_0: 基准医院的感染率
- $\beta_1, \beta_2, \beta_3$: 不同类型医院的感染率变化

例 1.5 Let $Y_i, i = 1, 2, \cdots, n$ be independent random variables from the above distribution with $\mu_i = \mathbb{E}[Y_i]$. Consider two explanatory variables x_1 and x_2 that are each measured on the ith observation. Write down a generalized linear model for these data and suggest one suitable choice of link function, explaining why it is appropriate. Give the model in vector-matrix form, making sure to specify the response vector $\boldsymbol{y}$, the parameter vector $\boldsymbol{\beta}$ and the design matrix $\boldsymbol{X}$.

解. 我们使用对数连接函数:

$$g(\mu_i) = \log(\mathbb{E}[Y]) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3.$$

用向量-矩阵形式表示模型:

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}.$$

$$\mu = \exp(\boldsymbol{X}\boldsymbol{\beta})$$

2 Logistic 回归

The table below shows results of a bioassay to compare the biological potencies of two preparations (batches) of insulin, by measuring the proportions of rodents that exhibit a particular response to different doses of each preparation.

研究人员进行了一项生物测定实验 *(bioassay)*, 以比较两批胰岛素 *(*标准批次 *vs* 试验批次*)* 的生物效能. 实验测量了不同剂量下动物产生反应的比例.

A different group of rodents is used for each dose of each preparation. It is desired to measure the potency of the test preparation relative to the standard where, for example, the potency is 2 if the proportion of responses produced by the standard preparation can be obtained using only half the dose of the test preparation.

Potency may be estimated by using logistic regression analysis, assuming that the rodents exhibit a logistic tolerance distribution in relationship to the $\log(\text{dose})$. In this analysis, $\log_{10}(\text{dose})$ is used as one explanatory variable and preparation (test or standard) as another. The following is abbreviated output from fitting two models, `m1` and `m0`, in R.

在该实验中, 每个剂量组对应一组新的动物, 避免了交叉干扰. 研究者希望估算试验批次相对于标准批次的效力 *(potency)*. 假设动物对剂量的对数 $(\log_{10}(dose))$ 呈现 *Logistic* 反应分布, 进行 *Logistic* 回归分析, 并通过广义线性模型 *(GLM)* 估算效力.

表 2: Data

Obs	Prep	Dose	Resp	Total
1	Standard	3.40	0	33
2	Standard	5.20	5	32
3	Standard	7.00	11	38
4	Standard	8.50	14	37
5	Standard	10.50	18	40
6	Standard	13.00	21	37
7	Standard	18.00	23	31
8	Standard	21.00	30	37
9	Standard	28.00	27	30
10	Test	6.50	2	40
11	Test	10.00	10	30
12	Test	14.00	18	40
13	Test	21.50	21	35
14	Test	29.00	27	37

```
> m1 <- glm(cbind(Resp,NonResp)~ log10(Dose)*
    Prep, family=binomial)
> summary(m1)

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.7907  0.6839 -8.467 <2e-16 ***
log10(Dose)  5.5180  0.6446  8.561 <2e-16 ***
PrepTest    -0.2170  1.2077 -0.180 0.857
log10(Dose):PrepTest -0.6269 1.0464 -0.599 0.549

Null deviance: 166.8335 on 13 degrees of freedom
Residual deviance: 8.4351 on 10 degrees of
    freedom
AIC: 64.287
```

```
> m0 <- glm(cbind(Resp,NonResp)~ log10(Dose)+
    Prep, family=binomial)
> summary(m0)

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.5531  0.5427 -10.23 < 2e-16 ***
log10(Dose)  5.2894  0.5057  10.46 < 2e-16 ***
PrepTest    -0.9290  0.2334  -3.98 6.89e-05 ***

Null deviance: 166.8335 on 13 degrees of freedom
Residual deviance: 8.7912 on 11 degrees of
    freedom
AIC: 62.644
```

Logistic 回归模型 用于建模二分类数据, 如成功/失败、存活/死亡. 模型形式为:

$$\text{logit } p(x) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots \quad \Longrightarrow \quad p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots)}}$$

2.1 交互项

例 2.1 Based on the above output, does the effect of preparation on the odds of the response depend on the dose? Explain.

效应是否依赖于计量?

解. 观察回归模型 m1 中交互项 log10(Dose) : PrepTest 的 p 值 $= 0.549 (> 0.05)$.

这说明交互项不显著, 意味着剂量效应在不同批次间没有显著差异.

2.2 模型选择

1. 观察模型在建模过程中的不同, 选择更简单且更有效的模型 (如: 交互项会提升模型复杂度, 其 p 值并不占优, 那么放弃这种建模方法);
2. 观察 AIC (Akaike information criterion), 越小越好.

解. 在两个模型中显然选择 m0, 理由是

1. 交互项 log10(Dose) : PrepTest 的 p 值 $= 0.549 (> 0.05)$, 这说明交互项不显著, 意味着剂量效应在不同批次间没有显著差异.
2. m0 的 AIC $= 62.644$ 比 m1 的 AIC $= 64.287$ 更低 (说明 m0 更好).

2.3 回归方程

根据上述结论选择 `m0` 后, 得到回归方程为:

$$\log\left(\frac{p}{1-p}\right) = -5.5531 + 5.2894 \cdot \log_{10}(\text{Dose}) - 0.9290 \cdot \text{PrepTest}$$

2.4 估算分位数

以中位数为例, 即令 $p = 0.5$, 那么

$$\log\left(\frac{0.5}{1-0.5}\right) = \log(1) = 0 = -5.5531 + 5.2894 \cdot \log_{10}(\text{Dose}) - 0.9290 \cdot \text{PrepTest}$$

$$5.2894 \cdot \log_{10}(\text{Dose}) = 5.5531 + 0.9290 \cdot \text{PrepTest}$$

$$\log_{10}(\text{Dose}) = \frac{5.5531 + 0.9290 \cdot \text{PrepTest}}{5.2894}$$

标准批次: 此时 $\text{PrepTest} = 0$, 使得

$$\log_{10}(\text{Dose}_{\text{Standard}}) = \frac{5.5531}{5.2894} = 1.04985$$

$$\text{Dose}_{\text{Standard}} = 10^{1.04985} = 11.21642$$

试验批次: 此时 PrepTest $= 1$, 使得

$$\log_{10}(\text{Dose}_{\text{Test}}) = \frac{5.5531 + 0.9290}{5.2894} = 1.22549$$
$$\text{Dose}_{\text{Test}} = 10^{1.22549} = 16.80694$$

2.5 效力 (Potency)

例 2.2 The ratio of the median effective doses of standard to test preparations is an estimate of the potency. Calculate the potency.

解. 由题易得, 效力公式为二者中位数之比, 即

$$\text{Potency} = \frac{\text{Median}(\text{Dose}_{\text{Standard}})}{\text{Median}(\text{Dose}_{\text{Test}})} = \frac{11.21642}{16.80694} = 0.66737$$

意味着标准批次的效力只占试验批次的 66.737%.

2.6 胜算比 (Odds Ratio, OR)

事件发生的概率与不发生的概率之比被称为胜算, 即

$$\text{Odds} = \frac{p}{1-p}.$$

若剂量翻倍, 则翻倍前后的胜算分别为

$$\begin{aligned}\text{Odds}_{\text{before}} &= \exp\Big(-5.5531 + 5.2894 \cdot \log_{10}(\text{Dose}) - 0.9290 \cdot \text{PrepTest}\Big)\\ \text{Odds}_{\text{after}} &= \exp\Big(-5.5531 + 5.2894 \cdot \log_{10}(\text{Dose} \times 2) - 0.9290 \cdot \text{PrepTest}\Big)\end{aligned}$$

因此, 胜算比为

$$\begin{aligned}\text{OR} = \frac{\text{Odds}_{\text{after}}}{\text{Odds}_{\text{before}}} &= \frac{\exp\Big(-5.5531 + 5.2894 \cdot \log_{10}(\text{Dose} \times 2) - 0.9290 \cdot \text{PrepTest}\Big)}{\exp\Big(-5.5531 + 5.2894 \cdot \log_{10}(\text{Dose}) - 0.9290 \cdot \text{PrepTest}\Big)}\\ &= \exp\Big(5.2894 \cdot \Big(\log_{10}(\text{Dose} \times 2) - \log_{10}(\text{Dose})\Big)\Big)\\ &= \exp\Big(5.2894 \cdot \log_{10} 2\Big)\\ &= 4.91488\end{aligned}$$

剂量翻倍, 成功概率约为原来的 4.91 倍.

2.7 模型的拟合优度 (Goodness of Fit)

2.7.1 残差偏差 (Residual Deviance)

- 残差偏差接近自由度 (Degrees of Freedom, df) 时, 说明模型拟合较好.

- 若残差偏差远大于 df, 说明模型拟合不好 (欠拟合) .
- 若残差偏差远小于 df, 说明模型可能过拟合 (过度拟合) .

根据 R output 中

```
> m1 <- glm(cbind(Resp,NonResp)~ log10(Dose)*
    Prep, family=binomial)
> summary(m1)

Residual deviance: 8.4351 on 10 degrees of
    freedom
AIC: 64.287
```

```
> m0 <- glm(cbind(Resp,NonResp)~ log10(Dose)+
    Prep, family=binomial)
> summary(m0)

Residual deviance: 8.7912 on 11 degrees of
    freedom
AIC: 62.644
```

即使看起来m1的 residual deviance 同m0比更接近 df, 但是代价是牺牲了 1 个自由度, 即模型更加复杂, 但是 $8.7912 - 8.4351 = 0.3561$, 这在统计上是微不足道的改进.

因此m0更优.

2.7.2 AIC

正如2.2中提及到的 m0 的 AIC $= 62.644$ 比 m1 的 AIC $= 64.287$ 更低 (说明 m0 更好).

2.7.3 Observed Values vs. Fitted Values

如表3所示, 显然m0的拟合值更接近真实值.

表 3: The fitted values

	Observed values	Fitted values	
		m1	m0
1	0	1.79	2.00
2	5	4.39	4.67
3	11	9.30	9.61
4	14	12.59	12.80
5	18	18.44	18.49
6	21	21.76	21.61
7	23	23.46	23.18
8	30	30.28	29.92
9	27	26.99	26.73
⋮	⋮	⋮	⋮

2.7.4 Hosmer-Lemeshow 拟合优度检验 (Optional)

若 Hosmer-Lemeshow p 值 > 0.05, 说明模型拟合良好

2.7.5 ROC (Receiver Operating Characteristic) 曲线 (Optional)

ROC 曲线下方的面积 (Area under the Curve (AUC) of ROC),

- AUC $= 1$, 是完美分类器, 采用这个预测模型时, 存在至少一个阈值能得出完美预测.
- $0.5 <$ AUC < 1, 优于随机猜测. 这个分类器 (模型) 妥善设置阈值的话, 能有预测价值.
- AUC $= 0.5$, 跟随机猜测一样 (例：丢铜板) , 模型没有预测价值.
- AUC < 0.5, 比随机猜测还差；但只要总是反预测而行, 就优于随机猜测.

3 Poisson 回归

3.1 泊松回归中的偏置项 (Offset)

- 泊松回归用于建模计数数据, 但在一些情况下, 数据的观测时间或规模不同, 需要进行标准化.
- 偏置项 (Offset) 允许我们调整不同观察单位的规模, 使得模型估计的是标准化的投诉率, 而非原始投诉数量.

比如, 不加偏置项的泊松回归:

$$\log(\mathbb{E}[Y_i]) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots$$

引入偏置项:

$$\log(\mathbb{E}[Y_i]) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \log(\text{visits}_i)$$

其中:

- Y_i: 第 i 个医生收到的投诉数.
- visits_i: 医生的就诊次数 (offset).

在不同医生就诊次数不均等的情况下, 使得比较投诉数时更公平.

3.2 R output

- `residency`, (R), a binary variable taking values N or Y corresponding to whether the doctor had completed residency training;
- `pay`, (P), giving the dollars per hour earned by the doctor;
- `hours`, (H), giving the total number of hours worked by the doctor that year.

```
glm(formula = complaints ~ residency + offset(log(visits)), family = poisson)

Coefficients:
     Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.4525  0.1026 -62.891 <2e-16 ***
residencyY  -0.3041  0.1725 -1.763  0.0779 .

   Null deviance: 63.435 on 43 degrees of freedom
Residual deviance: 60.245 on 42 degrees of freedom
AIC: 187.03

Number of Fisher Scoring iterations: 5
```

回归方程

$$\log(\mathbb{E}[Y]) = -6.4525 - 0.3041 \times \text{residencyY}$$

-0.3041 表示完成住院培训的医生的投诉数更少.

指数变换:

$$e^{-0.3041} \approx 0.74.$$

投诉数减少 26% (即 $1 - 0.74 = 0.26$) . 完成住院培训的医生的投诉率是未培训医生的 74%.

然而 $p = 0.0779$ (接近 0.05, 但不够显著) . 说明住院培训可能影响投诉率, 但统计显著性不强.

3.3 Fisher Scoring 迭代

两种拟合参数的迭代办法:

1. Newton-Raphson 方法:

$$\theta^{(t+1)} = \theta^{(t)} - \left[H(\theta^{(t)})\right]^{-1} \nabla L(\theta^{(t)})$$

2. Fisher Scoring 用期望信息矩阵代替 Hessian 矩阵, 提高稳定性:

$$\theta^{(t+1)} = \theta^{(t)} + I^{-1}(\theta^{(t)}) \nabla L(\theta^{(t)})$$

Fisher Scoring 迭代次数 = 5, 说明模型收敛较快.

用于最大似然估计参数, 替代 Newton-Raphson 提高计算稳定性.

3.4 最优模型选择

如表4所示.

表 4: A series of models was fitted to the number of complaints

Model	Deviance	差值	解释
Null	63.435		
H	57.347		
H + P	57.131	-0.216	几乎无改进
H + P + R	55.341	-1.79	R 变量有效
H + P + R + H*P	53.789	-1.552	H 和 P 存在交互作用
H + P + R + H*R	50.182	-3.607	H 和 R 存在交互作用
H + P + R + H*P + H*R	44.747	-5.435	** 最优 **
H + P + R + H*P + H*R + P*R	44.405	-0.342	P 和 R 的交互作用并不显著

3.5 过度离散 (Overdispersion)

泊松回归的一个重要假设是:

$$\mathrm{Var}\left[Y\right] = \mathbb{E}\left[Y\right]$$

但如果数据中方差**远大于**均值, 则称为过度离散 (Overdispersion). 出现这种情况的原因如下:

- 未包含关键解释变量: 可能有遗漏的影响因素.
- 数据有群组效应: 例如, 某些医生整体投诉率较高或较低.

过度离散的表现:

- 泊松模型的残差偏差（Residual Deviance）远大于自由度。
- AIC 值很高，表明模型拟合不佳。

如何处理办法:

1. 使用负二项回归 (Negative Binomial Regression):

$$\Pr(Y = y; \theta) = \binom{y + r - 1}{r - 1} \theta^r (1 - \theta)^y.$$

   负二项回归增加一个额外的离散参数, 允许方差大于均值:

$$\operatorname{Var}[Y] = \mathbb{E}[Y] + \alpha(\mathbb{E}[Y])^2$$

2. 调整标准误:
   过度离散可能来源于未考虑某些医生特征