Generalised Linear Models

Guanxu Gleason WANG

University of Glasgow

版本：Past Paper of STAT5019 Generalised Linear Models, UofG

更新：May 10, 2025

目录

1 指数族分布

1.1 Exponential Family of Distributions (指数族分布)

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

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

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 Thm.1.1中的结果计算数学期望得

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

定理 1.1 (Mean and variance) The random variable Y follows an exponential family distribution of the form $f(y;\theta) = \exp(a(y)b(\theta) + c(\theta) + d(y))$, can be expressed as

$$\mathbb{E}[a(Y)] = -\frac{c'(\theta)}{b'(\theta)}, \quad \text{and} \quad \operatorname{Var}[a(Y)] = \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{[b'(\theta)]^3}.$$

证明. First note that since $f(y;\theta)$ is a p.d.f.

$$\int f(y;\theta)\,\mathrm{d}y = 1 \quad \Rightarrow \quad \frac{\mathrm{d}}{\mathrm{d}\theta}\int f(y;\theta)\,\mathrm{d}y = 0 \quad \Rightarrow \quad \int \frac{\mathrm{d}}{\mathrm{d}\theta} f(y;\theta)\,\mathrm{d}y = 0.$$

For the derivative,

$$\frac{\mathrm{d}}{\mathrm{d}\theta} f(y;\theta) = [a(y)b'(\theta) + c'(\theta)]f(y;\theta).$$

Integrating with respect to y then gives

$$\begin{aligned}
\int \frac{\mathrm{d}}{\mathrm{d}\theta} f(y;\theta)\,\mathrm{d}y &= \int [a(y)b'(\theta) + c'(\theta)]f(y;\theta)\,\mathrm{d}y \\
&= b'(\theta)\int a(y)f(y;\theta)\,\mathrm{d}y + c'(\theta)\int f(y;\theta)\,\mathrm{d}y \\
&= b'(\theta)\mathbb{E}[a(Y)] + c'(\theta) = 0 \quad \Rightarrow \quad \mathbb{E}[a(Y)] = -\frac{c'(\theta)}{b'(\theta)}.
\end{aligned}$$

Also, taking the second derivative gives

$$\begin{aligned}\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}f(y;\theta) &= [a(y)b''(\theta)+c''(\theta)]f(y;\theta)+[a(y)b'(\theta)+c'(\theta)]^2 f(y;\theta)\\ &= [a(y)b''(\theta)+c''(\theta)]f(y;\theta)+[b'(\theta)]^2\left[a(y)+\frac{c'(\theta)}{b'(\theta)}\right]^2 f(y;\theta)\\ &= [a(y)b''(\theta)+c''(\theta)]f(y;\theta)+[b'(\theta)]^2\left[a(y)-\mathbb{E}\left[a(Y)\right]\right]^2 f(y;\theta).\end{aligned}$$

Integrating with respect to y then gives

$$\begin{aligned}&\int \frac{\mathrm{d}^2}{\mathrm{d}\theta^2}f(y;\theta)\,\mathrm{d}y = b''(\theta)\mathbb{E}\left[a(Y)\right]+c''(\theta)+[b'(\theta)]^2\mathrm{Var}\left[a(Y)\right]=0\\ \Rightarrow\quad &\mathrm{Var}\left[a(Y)\right]=\frac{b''(\theta)\mathbb{E}\left[a(Y)\right]+c''(\theta)}{[b'(\theta)]^2}=\frac{b''(\theta)c'(\theta)-c''(\theta)b'(\theta)}{[b'(\theta)]^3}.\end{aligned}$$

定义 1.1 (Score statistic) The derivative of the log-likelihood $\ell(\theta;y)$ with respect to the parameter θ, i.e.

$$U=\frac{\mathrm{d}}{\mathrm{d}\theta}\ell(\theta;y)=\frac{\mathrm{d}}{\mathrm{d}\theta}\log\mathcal{L}(\theta;y).$$

对于指数族分布来说, 其得分为

$$U(\theta;y)=\frac{\mathrm{d}}{\mathrm{d}\theta}\ell(\theta;y)=a(y)b'(\theta)+c'(\theta).$$

因此

$$\mathbb{E}[U] = b'(\theta)\mathbb{E}[a(y)] + c'(\theta) = b'(\theta)\left(-\frac{c'(\theta)}{b'(\theta)}\right) + c'(\theta) = 0, \quad \text{and} \quad \operatorname{Var}[U] = [b'(\theta)]^2 \operatorname{Var}[a(Y)].$$

注意在求导的时候使用链式法则, 即若 $p = p(\theta)$, 则 $U(\theta; y) = \dfrac{\mathrm{d}}{\mathrm{d}\theta}\ell(p(\theta); y) = \dfrac{\mathrm{d}\ell}{\mathrm{d}p} \times \dfrac{\mathrm{d}p}{\mathrm{d}\theta}$.

定义 1.2 (Fisher's information) The **Fisher's information**, denoted as $\mathcal{I}$, is given by

$$\mathcal{I} = \operatorname{Var}[U] = \mathbb{E}\left[U^2\right] = \mathbb{E}\left[\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\ell(\theta; y)\right)^2\right] = -\mathbb{E}\left[\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\ell(\theta; y)\right] = -\mathbb{E}[U'].$$

For members of the exponential family of distributions, since

$$\operatorname{Var}[a(Y)] = \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{[b'(\theta)]^3},$$

we have that

$$\mathcal{I} = \operatorname{Var}[U] = [b'(\theta)]^2 \operatorname{Var}[a(Y)] = \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{b'(\theta)} = \frac{b''(\theta)c'(\theta)}{b'(\theta)} - c''(\theta).$$

1.2 Generalised Linear Models

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.

定义 1.3 (Generalised linear models) Let Y_i be independent responses from an exponential family distribution in *canonical form* ($a(y) = y$) and $\mu_i = \mathbb{E}[Y_i]$ for $i = 1, 2, \cdots, n$.

A **generalised linear model (GLM)** is a model of the form $g(\mu_i) = x_i^\top \boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is a p-dimensional parameter vector, $x_i^\top$ is the ith row of the design matrix $\boldsymbol{X}$, and $g(\cdot)$ is a monotonic (单调), differentiable function called the **link function**, a.k.a. **canonical link**.

表 1: 常见数据类型与分布选择及链接函数

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

1.3 Maximum Likelihood Estimation of GLM Coefficients

我们此前提到 score, $U(\theta) = \dfrac{\mathrm{d}}{\mathrm{d}\theta}\ell(\theta; y)$, 很显然, $U(\theta) = 0$ 的根就是 $\widehat{\theta}_{\text{MLE}}$, 即参数的最大似然估计.

当方程 $U(\theta) = 0$ 不易求解时, 那么有两种拟合参数的迭代办法:

1. Newton-Raphson 方法:

$$\theta^{(t+1)} = \theta^{(t)} - \frac{U^{(t)}}{U'^{(t)}}.$$

2. Fisher scoring: 用 U' 的数学期望 $\mathbb{E}[U'] = -\mathcal{I}$, 即 Fisher's information, 代替 U', 从而提高稳定性:

$$\theta^{(t+1)} = \theta^{(t)} + \frac{U^{(t)}}{\mathcal{I}^{(t)}}.$$

1.4 Deviance

定义 1.4 (Deviance) The deviance, D, is defined as

$$D = 2\log\lambda = 2\left[\ell\left(\widehat{\boldsymbol{\beta}}_{\max}; \boldsymbol{y}\right) - \ell\left(\widehat{\boldsymbol{\beta}}; \boldsymbol{y}\right)\right] \sim \chi^2(m-p),$$

where $\ell\left(\widehat{\boldsymbol{\beta}}_{\max}; \boldsymbol{y}\right)$ is the maximised log-likelihood for the saturated model (饱和模型) and $\ell\left(\widehat{\boldsymbol{\beta}}; \boldsymbol{y}\right)$ is the maximised log-likelihood for the model of interest.

The likelihood ratio $\lambda = \dfrac{\mathcal{L}\left(\widehat{\boldsymbol{\beta}}_{\max}; \boldsymbol{y}\right)}{\mathcal{L}\left(\widehat{\boldsymbol{\beta}}; \boldsymbol{y}\right)}$ provides a measure of how well the model of interest fits compared with the full model. 也就是说, $\mathcal{L}\left(\widehat{\boldsymbol{\beta}}; \boldsymbol{y}\right)$ 就是本身的似然函数, 而 full model 是把参数替换为 y_i.

In practice we often use the logarithm of the likelihood ratio:

$$\log\lambda = \log\mathcal{L}\left(\widehat{\boldsymbol{\beta}}_{\max}; \boldsymbol{y}\right) - \log\mathcal{L}\left(\widehat{\boldsymbol{\beta}}; \boldsymbol{y}\right) = \ell\left(\widehat{\boldsymbol{\beta}}_{\max}; \boldsymbol{y}\right) - \ell\left(\widehat{\boldsymbol{\beta}}; \boldsymbol{y}\right).$$

Hypothesis testing Compare nested models M_0 and M_1 using the difference of their deviances.

Consider

- $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0 = (\beta_1, \beta_2, \cdots, \beta_q)^\top$ corresponding to M_0; v.s.
- $H_1 : \boldsymbol{\beta} = \boldsymbol{\beta}_1 = (\beta_1, \beta_2, \cdots, \beta_p)^\top$ corresponding to M_1, with $q < p < n$.

Test H_0 against H_1 by considering

$$\begin{aligned} D_0 - D_1 &= 2\left[\ell\left(\widehat{\boldsymbol{\beta}}_{\max}; \boldsymbol{y}\right) - \ell\left(\widehat{\boldsymbol{\beta}}_0; \boldsymbol{y}\right)\right] - 2\left[\ell\left(\widehat{\boldsymbol{\beta}}_{\max}; \boldsymbol{y}\right) - \ell\left(\widehat{\boldsymbol{\beta}}_1; \boldsymbol{y}\right)\right] \\ &= 2\left[\ell\left(\widehat{\boldsymbol{\beta}}_1; \boldsymbol{y}\right) - \ell\left(\widehat{\boldsymbol{\beta}}_0; \boldsymbol{y}\right)\right] \sim \chi^2(p-q). \end{aligned}$$

If both models describe the data well then $D_0 \sim \chi^2(n-q),\ D_1 \sim \chi^2(n-p)$, and $D_0 - D_1 \sim \chi^2(p-q)$.

证明. Recall the Taylor series expansion for the log-likelihood

$$\ell(\boldsymbol{\beta}) \approx \ell(\widehat{\boldsymbol{\beta}}) + (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})^\top U(\widehat{\boldsymbol{\beta}}) - \frac{1}{2}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})^\top \mathcal{I}(\widehat{\boldsymbol{\beta}})(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}).$$

Since $\widehat{\boldsymbol{\beta}}$ is the MLE, $U(\widehat{\boldsymbol{\beta}}) = \mathbf{0}$ and hence this becomes

$$\ell(\boldsymbol{\beta}) - \ell(\widehat{\boldsymbol{\beta}}) \approx -\frac{1}{2}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})^\top \mathcal{I}(\widehat{\boldsymbol{\beta}})(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}).$$

Therefore, the statistic

$$2\left[\ell(\widehat{\boldsymbol{\beta}}; \boldsymbol{y}) - \ell(\boldsymbol{\beta}; \boldsymbol{y})\right] \approx (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})^\top \mathcal{I}(\widehat{\boldsymbol{\beta}})(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})$$

will be approximately $\chi^2(p)$. (底部有证明)

2 Models for binary/binomial response – 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, 即 $Y_i \overset{\text{i.i.d.}}{\sim} \text{Binomial}(n_i, p_i)$, 如成功/失败、存活/死亡等.

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

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

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

标准批次: 此时 $\text{PrepTest} = 0$, 使得

$$\begin{aligned}
\log_{10}(\text{Dose}_{\text{Standard}}) &= \frac{5.5531}{5.2894} = 1.04985, \\
\text{Dose}_{\text{Standard}} &= 10^{1.04985} = 11.21642.
\end{aligned}$$

试验批次: 此时 $\text{PrepTest} = 1$, 使得

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

↪ **超纲内容: 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%.

2.5 胜算比 (Odds Ratio, OR)

定义 2.1 (Odd) 事件发生的概率与不发生的概率之比被称为**胜算 (odd)**, 即

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

In logistic regression we model the **log odds**:

$$\log(\text{Odds}) = \log\left(\frac{p}{1-p}\right) = \boldsymbol{x}^\top \boldsymbol{\beta}.$$

The β coefficients are **log odds ratios**:

$$\beta = \log(\text{Odds}_1) - \log(\text{Odds}_2) = \log\left(\frac{\text{Odds}_1}{\text{Odds}_2}\right) = \log\left(\frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}}\right).$$

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.

2.6 模型的拟合优度 (Goodness of Fit)

2.6.1 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.

2.6.2 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).

2.6.3 Observed Values vs. Fitted Values

如表3所示, 显然`m0`的拟合值更接近真实值.

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

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

2.6.5 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.

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
⋮	⋮	⋮	⋮

3 Models for count responses – Poisson 回归

For the number of occurrences Y, we assume that Y follows the Poisson distribution $\text{Poisson}(\mu)$ with probability mass function given by

$$f(y) = \frac{\mu^y e^{-\mu}}{y!}.$$

The mean and variance of $Y \sim \text{Poisson}(\mu)$ are both equal to μ.

3.1 泊松回归中的偏置项 (Offset)

Let $Y_1, Y_2, \cdots, Y_n$ be independent random variables with Y_i denoting the number of events occurred from exposure n_i for the ith covariate pattern. Then

$$\mathbb{E}[Y_i] = \mu_i = n_i \theta_i.$$

The dependence on explanatory variables is usually modelled by $\theta_i = e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}$. The corresponding GLM is

$$\mathbb{E}[Y_i] = \mu_i = n_i e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}; \quad Y_i \sim \text{Poisson}(\mu_i).$$

This corresponds to the log link:

$$\log \mu_i = \log n_i + \boldsymbol{x}_i^\top \boldsymbol{\beta},$$

where the term $\log n_i$ is called the **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

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

其中, -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 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

3.4 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

4 考试大纲

Part 1 – 指数族分布

- 指数族分布的性质推导;
- 公式（20）的 score 推导和公式（21）的 information in the general case 推导. 可能会要求你获取特定分布、模型和链接函数的得分和信息.
  - 公式（20）的 score 推导:

    此前提及过在计算 score statistic 时，使用链式法则，当我们想用 OLS 估计 β_j 时，

$$U_j = \frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} \left[\frac{\partial \ell_i}{\partial \beta_j} \right] = \sum_{i=1}^{n} \left[\frac{\partial \ell_i}{\partial \theta_i} \cdot \frac{\partial \theta_i}{\partial \mu_i} \cdot \frac{\partial \mu_i}{\partial \beta_j} \right],$$

    其中

$$\begin{aligned}
\frac{\partial \ell_i}{\partial \theta_i} &= y_i b'(\theta) + c'(\theta) = b'(\theta) \left[y_i + \frac{c'(\theta)}{b'(\theta)} \right] = b'(\theta)(y_i - \mu_i), \\
\frac{\partial \mu_i}{\partial \theta_i} &= -\frac{c''(\theta_i) b'(\theta_i) - c'(\theta_i) b''(\theta_i)}{[b'(\theta_i)]^2} = b'(\theta_i) \mathrm{Var}\left[Y_i\right] \\
&\Rightarrow \frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{b'(\theta_i) \mathrm{Var}\left[Y_i\right]}, \\
\frac{\partial \mu_i}{\partial \beta_j} &= \frac{\partial \mu_i}{\partial g(\mu_i)} \cdot \frac{\partial g(\mu_i)}{\partial \beta_j} = \frac{1}{g'(\mu_i)} \cdot \frac{\partial (\boldsymbol{x}_i^\top \boldsymbol{\beta})}{\partial \beta_j} = \frac{x_{ij}}{g'(\mu_i)}.
\end{aligned}$$

因此

$$U_j = \sum_{i=1}^{n} \left[b'(\theta)(y_i - \mu_i) \cdot \frac{1}{b'(\theta_i)\mathrm{Var}\left[Y_i\right]} \cdot \frac{x_{ij}}{g'(\mu_i)} \right] = \sum_{i=1}^{n} \left[\frac{y_i - \mu_i}{\mathrm{Var}\left[Y_i\right]} \cdot \frac{x_{ij}}{g'(\mu_i)} \right]. \tag{Ch1.20}$$

- 公式（21）的 information in the general case 推导:
  因为 $\mathcal{I} = \mathrm{Var}\left[U\right]$, 因此当 U 时一个矩阵时, 则

$$\begin{aligned}
\mathcal{I}_{jk} &= \mathrm{Cov}\left[U_j, U_k\right] = \mathbb{E}\left[U_j U_k\right] - \mathbb{E}\left[U_j\right]\mathbb{E}\left[U_k\right] \\
&= \mathbb{E}\left[U_j U_k\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{n} \left[\frac{y_i - \mu_i}{\mathrm{Var}\left[Y_i\right]} \cdot \frac{x_{ij}}{g'(\mu_i)} \right] \sum_{i=1}^{n} \left[\frac{y_i - \mu_i}{\mathrm{Var}\left[Y_i\right]} \cdot \frac{x_{ik}}{g'(\mu_i)} \right] \right] \\
&= \sum_{i=1}^{n} \left[\frac{\mathbb{E}\left[(Y_i - \mu_i)^2\right]}{(\mathrm{Var}\left[Y_i\right])^2} \cdot \frac{x_{ij}x_{ik}}{[g'(\mu_i)]^2} \right] \\
&= \frac{x_{ij}x_{ik}}{\mathrm{Var}\left[Y_i\right] \cdot [g'(\mu_i)]^2}.
\end{aligned} \tag{Ch1.21}$$

Notice that the information matrix can be written as

$$\mathcal{I} = \mathcal{I}(\boldsymbol{\beta}) = \boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X},$$

where $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1^\top \\ \boldsymbol{x}_2^\top \\ \vdots \\ \boldsymbol{x}_n^\top \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$, $\boldsymbol{W} = \operatorname{diag}(\boldsymbol{w}) = \begin{pmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & w_n \end{pmatrix}$, and

$$w_i = \frac{1}{\operatorname{Var}[Y_i] \cdot [g'(\mu_i)]^2}, \quad i = 1, 2, \cdots, n.$$

因此, 公式 (20) 可以写作

$$U_j = \sum_{i=1}^{n} \left[\frac{y_i - \mu_i}{\operatorname{Var}[Y_i]} \cdot \frac{x_{ij}}{g'(\mu_i)} \right] = \sum_{i=1}^{n} \left[\frac{(y_i - \mu_i) \cdot x_{ij} \cdot g'(\mu_i)}{\operatorname{Var}[Y_i] \cdot [g'(\mu_i)]^2} \right] = \sum_{i=1}^{n} x_{ij} w_i z_i,$$

where $z_i = (y_i - \mu_i) \cdot g'(\mu_i)$, so the score can be expressed in vector-matrix form as

$$\boldsymbol{U}(\boldsymbol{\theta}) = \boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{z}.$$

故, the Fisher's method of scoring is based on the estimating equation

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}^{(m)} &= \widehat{\boldsymbol{\beta}}^{(m-1)} + \left[\mathcal{I}^{(m-1)}\right]^{-1} \boldsymbol{U}^{(m-1)} \\
&= \widehat{\boldsymbol{\beta}}^{(m-1)} + \left[\boldsymbol{X}^\top \boldsymbol{W}^{(m-1)} \boldsymbol{X}\right]^{-1} \boldsymbol{X}^\top \boldsymbol{W}^{(m-1)} \boldsymbol{z}^{(m-1)} \\
&= \left(\left[\boldsymbol{X}^\top \boldsymbol{W}^{(m-1)} \boldsymbol{X}\right]^{-1} \boldsymbol{X}^\top \boldsymbol{W}^{(m-1)} \boldsymbol{X}\right) \widehat{\boldsymbol{\beta}}^{(m-1)} + \left[\boldsymbol{X}^\top \boldsymbol{W}^{(m-1)} \boldsymbol{X}\right]^{-1} \boldsymbol{X}^\top \boldsymbol{W}^{(m-1)} \boldsymbol{z}^{(m-1)} \\
&= \left[\boldsymbol{X}^\top \boldsymbol{W}^{(m-1)} \boldsymbol{X}\right]^{-1} \boldsymbol{X}^\top \boldsymbol{W}^{(m-1)} \left(\boldsymbol{X} \widehat{\boldsymbol{\beta}}^{(m-1)} + \boldsymbol{z}^{(m-1)}\right) \\
&= \left[\boldsymbol{X}^\top \boldsymbol{W}^{(m-1)} \boldsymbol{X}\right]^{-1} \boldsymbol{X}^\top \boldsymbol{W}^{(m-1)} \left(\boldsymbol{\eta}^{(m-1)} + \boldsymbol{z}^{(m-1)}\right). \qquad \text{(Ch1.27)}
\end{aligned}$$

- 导数（涉及泰勒级数展开）用于样本分布的得分、最大似然估计和偏差。需要知道分布结果。
  - Inference.

    上面提到 $\mathcal{I}_{jk} = \text{Cov}\left[U_j, U_k\right] = \mathbb{E}\left[U_j U_k\right]$, 因此对于 univariate β, 有

$$\frac{U}{\sqrt{\mathcal{I}}} \overset{\text{approx}}{\sim} \mathcal{N}(0,1) \quad \Leftrightarrow \quad \frac{U^2}{\mathcal{I}} \overset{\text{approx}}{\sim} \chi^2(1).$$

    那么, 对于 $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)^\top$, 则有

$$\boldsymbol{U} \overset{\text{approx}}{\sim} \mathcal{N}_p(0, \mathcal{I}) \quad \Rightarrow \quad \boldsymbol{U}^\top \mathcal{I}^{-1} \boldsymbol{U} \overset{\text{approx}}{\sim} \chi^2(p).$$

  - 泰勒级数展开的应用.

$$\ell(\beta) \approx \ell(\widehat{\beta}) + (\beta - \widehat{\beta}) \left.\frac{\partial \ell}{\partial \beta}\right|_{\beta=\widehat{\beta}} + \frac{1}{2}(\beta - \widehat{\beta})^2 \left.\frac{\partial^2 \ell}{\partial \beta^2}\right|_{\beta=\widehat{\beta}}.$$

由于 $\mathcal{I} = -\mathbb{E}\left[\dfrac{\mathrm{d}^2}{\mathrm{d}\beta^2}\ell(\beta; y)\right]$, 则

$$\ell(\beta) \approx \ell(\widehat{\beta}) + (\beta - \widehat{\beta})U(\widehat{\beta}) - \frac{1}{2}(\beta - \widehat{\beta})^2\mathcal{I}(\widehat{\beta}).$$

For a vector $\boldsymbol{\beta}$ of parameters this generalises to

$$\ell(\boldsymbol{\beta}) \approx \ell(\widehat{\boldsymbol{\beta}}) + (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})^\top U(\widehat{\boldsymbol{\beta}}) - \frac{1}{2}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})^\top\mathcal{I}(\widehat{\boldsymbol{\beta}})(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}).$$

Similarly for the score function with a single parameter β the Taylor series expansion is

$$U(\beta) \approx U(\widehat{\beta}) + (\beta - \widehat{\beta})U'(\widehat{\beta}) = U(\widehat{\beta}) - (\beta - \widehat{\beta})\mathcal{I}(\widehat{\beta}).$$

For a vector $\boldsymbol{\beta}$ of parameters this generalises to

$$U(\boldsymbol{\beta}) \approx U(\widehat{\boldsymbol{\beta}}) - \mathcal{I}(\widehat{\boldsymbol{\beta}})(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}).$$

如果 $\boldsymbol{\beta}$ 是通过 MLE 估计的, 那么 $U(\widehat{\boldsymbol{\beta}}) = 0$, 因此

$$U(\boldsymbol{\beta}) \approx -\mathcal{I}(\widehat{\boldsymbol{\beta}})(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) \quad \Rightarrow \quad \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \mathcal{I}^{-1}\boldsymbol{U}.$$

The variance-covariance matrix for $\boldsymbol{\beta}$ is

$$\mathbb{E}\left[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top\right] = \mathbb{E}\left[\mathcal{I}^{-1}\boldsymbol{U}\boldsymbol{U}^\top\mathcal{I}^{-\top}\right] = \mathcal{I}^{-1}\mathbb{E}\left[\boldsymbol{U}\boldsymbol{U}^\top\right]\mathcal{I}^{-\top} = \mathcal{I}^{-1}.$$

因此

$$(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top\mathcal{I}(\widehat{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \overset{\text{approx}}{\sim} \chi^2(p) \quad \Leftrightarrow \quad \widehat{\boldsymbol{\beta}} \overset{\text{approx}}{\sim} \mathcal{N}_p(\boldsymbol{\beta}, \mathcal{I}^{-1}).$$

- 关于 hat matrix 的详细信息，标准化残差以及检查异常值 (outliers)/影响点 (influential observations) 的办法。

  For a normal linear model, the residuals are given by $\widehat{e}_i = y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}} = y_i - \widehat{\mu}_i$. The variance-covariance matrix of the vector of residuals $\widehat{e}_i$ is given by

$$\begin{aligned}
\mathbb{E}\left[\widehat{\boldsymbol{e}}\widehat{\boldsymbol{e}}^\top\right] &= \mathbb{E}\left[(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^\top\right] \\
&= \mathbb{E}\left[\boldsymbol{y}\boldsymbol{y}^\top\right] - \boldsymbol{X}\mathbb{E}\left[\widehat{\boldsymbol{\beta}}\widehat{\boldsymbol{\beta}}^\top\right]\boldsymbol{X}^\top \\
&= \sigma^2\left[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\right] \\
&= \sigma^2\left[\boldsymbol{I} - \boldsymbol{H}\right],
\end{aligned}$$

  where $\boldsymbol{I}$ is the identity matrix and $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$ is the **hat matrix**.

  We use the **standardised residuals**

$$r_i = \frac{\widehat{e}_i}{\widehat{\sigma}(1 - h_{ii})^{1/2}}$$

  to check the adequacy of the model. Under the **model assumptions** the standardised residuals are **Outliers**

	Model assumption	Diagnostic plot
1	Linearity of relationships between variables	Standardised residuals against each of the explanatory variables
2	Normality of errors	Normal probability plot (residuals against their expected values) standardised residuals against the fitted values $\widehat{y}_i$
3	Constant variance	
4	Serial independence of observations	Sequence plot of the standardised residuals

are observations that are not well fitted by the model.

Influential observations have a large effect on inferences based on the model. Influential observations may or may not be outliers and vice versa.

The ith diagonal element of the hat matrix, h_{ii}, is known as the **leverage** of the ith observation.

Part 2 – Logistic 回归

- 偏差残差 (deviance residuals) 公式.

定义 4.1 (Pearson's chi-squared statistic)

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - n_i\widehat{p}_i)^2}{n_i\widehat{p}_i(1-\widehat{p}_i)} \sim \chi^2(n-p), \quad i = 1, 2, \cdots, n,$$

where y_i represents the observed number of successes, n_i is the number of trials, and $\widehat{p}_i$ for the ith covariate pattern. For $\chi^2(n-p)$, n is the number of parameters in the saturated model (usually equal to the number of observations), and p is the number of parameters in the model of interest.

定义 4.2 (Pearson residuals)

$$X_k = \frac{y_k - n_k\widehat{p}_k}{\sqrt{n_k\widehat{p}_k(1-\widehat{p}_k)}}.$$

The **standardised Pearson residual** is

$$r_{Pk} = \frac{X_k}{\sqrt{1-h_k}},$$

where h_k is the leverage which is obtained from the hat matrix.

定义 4.3 (Deviance residuals)

$$d_k = \text{sign}(y_k - n_k\widehat{p}_k) \times \left\{ 2\left[y_k \log\left(\frac{y_k}{n_k\widehat{p}_k}\right) + (n_k - y_k)\log\left(\frac{n_k - y_k}{n_k - n_k\widehat{p}_k}\right)\right]\right\}^{\frac{1}{2}}.$$

The **standardised deviance residual** is

$$r_{Dk} = \frac{d_k}{\sqrt{1-h_k}},$$

where h_k is the leverage which is obtained from the hat matrix.

- 有序逻辑回归模型 (ordinal logistic regression models).
  存在一个**潜在变量 (latent variable)**, Z, 其中存在 $J-1$ 个**切割点 (cutpoints)** 分别为 $C_1, C_2, \cdots, C_{J-1}$, 将 Z 分割成 J 个区间分别为 $p_1 p_2, \cdots, p_J$. 常用的是**比例风险逻辑回归模型 (proportional odds logistic regression)**.

定义 4.4 (Proportional odds logistic regression) 如果一个线性预测器 $\boldsymbol{x}^\top\boldsymbol{\beta}_j$ 截距项 β_{0j} 依赖第 j 个类别 (即分割后的区间), 但是其他的解释变量并不依赖 j, 则模型为

$$\log\left(\frac{p_1 + p_2 + \cdots + p_j}{p_{j+1} + p_{j+2} + \cdots + p_J}\right) = \beta_{0j} + \beta_1 x_1 + \cdots + \beta_{p-1}x_{p-1}.$$

替代方案:

- Cumulative (累计) logit model:

$$\log\left(\frac{\Pr(Z \leqslant C_j)}{\Pr(Z > C_j)}\right) = \log\left(\frac{p_1 + p_2 + \cdots + p_j}{p_{j+1} + p_{j+2} + \cdots + p_J}\right) = \boldsymbol{x}^\top\boldsymbol{\beta}_j.$$

- Adjacent (相邻) categories logit model:

$$\log\left(\frac{p_j}{p_{j+1}}\right) = \boldsymbol{x}^\top \boldsymbol{\beta}_j.$$

- Continuation ratio (续比) logit model:
  比如 $\frac{p_1}{p_2}, \frac{p_1+p_2}{p_3}, \cdots, \frac{p_1+p_{J-1}}{p_J}$, 或者 $\frac{p_1}{p_2+\cdots+p_J}, \frac{p_2}{p_3+\cdots+p_J}, \cdots, \frac{p_{J-1}}{p_J}$.

Part 3 – Poisson 回归

拟合零膨胀 (zero-inflated) 和门槛 (hurdle) 模型. 足够熟悉它们何时适用.

数据中出现大量的 0 时适用.

案例 1 (Zero-inflated Models) `zeroinfl()`

在零膨胀泊松或负二项式模型中，我们假设有两个可能产生数据中零的过程。一个是伯努利过程，另一个是泊松过程，结果的数据分布是这两个过程的混合。此类模型可能在以下情境中适用:

- 物品购买次数: 有些人可能不会买任何东西 (因个人意愿), 即使有意愿买也可能因为缺货等外部原因最终为 0.
- 看病次数: 受性别/年龄等影响, 有些人从不去看病 (结构性零), 有些人会但在某段时间没去 (非结构性零)。

案例 2 (Hurdle Models) `hurdle()`

它假设存在一个顺序过程, 分为以下两个步骤进行建模:

1. 先用一个模型判断结果是否为 0 (是否跨过“门槛”);
2. 若结果非 0, 则用另一个分布 (通常是截断的泊松或负二项) 对正数部分建模.

比如:

- 住院天数: 病人要么没住院（0），要么住了若干天（>0），不会出现负数;
- 吸烟数量: 有些人根本不吸烟（结构性零），吸烟者会有非零的消费量。

特点	Zero-inflated Model	Hurdle Model
零的来源	来自两种过程（伯努利 + 泊松/负二项）	来自一步决策过程（是否跨过门槛）
正值建模	包括 0 和正数	只对正数建模（0 被单独处理）
模型结构	混合模型（两个过程叠加）	阶段性模型（先判断 0，再建模非零部分）
常用于	结构性和非结构性 0 混合场景	0 和正值完全分离的情境