Análise de Dados Categorizados - Aula 9

Márcia D Elia Branco

Universidade de São Paulo
Instituto de Matemática e Estatística
mdbranco@usp.br - sala 295-A -

Exemplo 2: Os dados no arquivo *Crabs.dat* (Agresti, 2019) referem-se a um estudo de aninhamento de caranguejos-ferradura. Neste estudo, foram contados quantos machos satélites estão associados a cada fêmea do caranguejo-ferradura. Deseja-se investigar a associação entre o comprimento da carapaça do caranguejo com o número satélites. O tamanho de amostra é $n = 173$. Na tabela abaixo apresentamos um subconjunto desses dados

Satélites	8	4	0	0	0	0
Tamanho	28.3	26.0	25.6	21.0	22.5	23.8
Satélites	0	14	9	0	0	8
Tamanho	24.3	26.0	26.0	24.7	25.8	27.1

MLG: modelo log linear Poisson

MLG: modelo log linear Poisson

- A variável resposta Y é o número de machos satélites. A variável explicativa x é o tamanho.
- Suposição $Y \sim Poisson(\mu)$.
- A função de ligação considerada é $g(\mu) = \log(\mu)$
- O preditor linear é $\alpha + \beta x$
- A função no programa **R** que ajusta MLG é

$$glm(y \sim x, family = poisson(link = log), data = Crabs)$$

MLG: modelo log linear Poisson

O modelo ajustado:

$$\log(\hat{\mu}) = -3.3 + 0.164x$$

Portanto, para obtenção dos valores ajustados pelo modelo para as médias calculamos

$$\hat{\mu} = \exp\{-3.3 + 0.164x\}.$$

Por exemplo, para uma fêmea de comprimento 26 cm estima-se uma média de 2.6 machos satélites.

MLG: modelo log linear Poisson

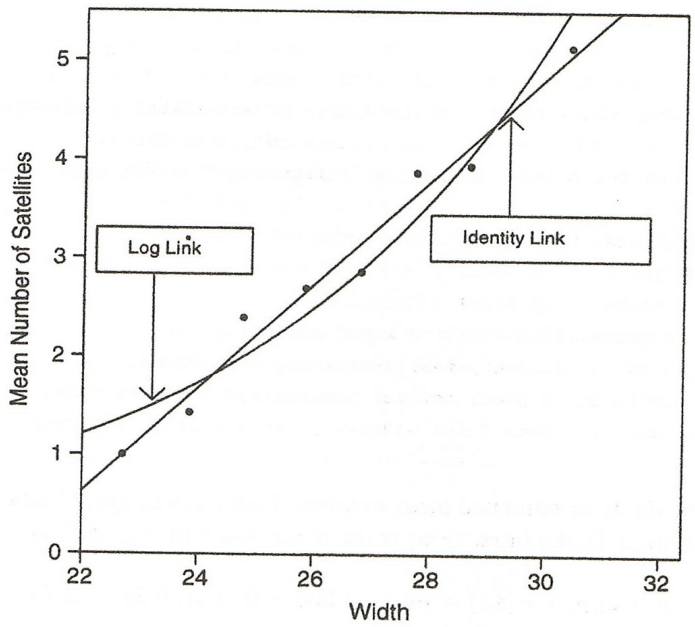

Figure 3.4 Estimated mean number of satellites for log and identity links.

MLG: modelo log linear Poisson

Teste de hipóteses:

$$H_0 : \beta = 0 \quad \textit{versus} \quad H_a : \beta \neq 0$$

Estatística de Wald:

$$z^2 = \frac{(0.164)^2}{(0.01997)^2} = 67.44$$

$Valor - P < 10^{-15}$. Rejeita-se H_0.

O comprimento da carapaça é uma variável significativa para o modelo.

MLG estimação dos parâmetros

Para o caso geral de MLG, temos que a log-verossimilhança é

$$\log(L(\theta)) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)$$

Considere um preditor linear $\eta_i = \beta x_i$, em que $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ é um vetor de parâmetros e $x_i = (1, x_{i1}, \ldots, x_{ip})$.
Além disso,

$$g(\mu_i) = \eta_i = \beta x_i \quad , \quad i = 1, \ldots, n.$$

Para obtenção do emv dos coeficientes regressores, precisamos derivar a log-verossimilhança em cada componente do vetor β.

Denotamos por L_i o i-ésimo componente da soma na expressão da função de verosimilhança.

MLG estimação dos parâmetros

Pela regra da cadeia, temos que

$$\frac{d \log L_i}{d\beta_j} = \left(\frac{d \log L_i}{d\theta_i}\right)\left(\frac{d\theta_i}{d\mu_i}\right)\left(\frac{d\mu_i}{d\eta_i}\right)\left(\frac{d\eta_i}{d\beta_j}\right)$$

Vamos usar as seguintes propriedades da família exponencial:

$$\mu_i = E[Y_i] = \frac{db(\theta_i)}{d\theta_i} \quad \text{e} \quad a(\phi)Var(Y_i) = \frac{d^2 b(\theta_i)}{d\theta_i^2}$$

Então,

$$\frac{d \log L_i}{d\theta_i} = \frac{y_i - \mu_i}{a(\phi)}$$

e

$$\frac{d\theta_i}{d\mu_i} = \frac{a(\phi)}{Var(Y_i)}$$

Além disso, $\frac{d\eta_i}{d\beta_j} = x_{ij}$.

Finalmente, notamos que $\frac{d\mu_i}{d\eta_i}$ depende da escolha da função de ligação.

Assim, as equações de estimação são dadas por

$$\frac{d \log L(\beta)}{d\beta_j} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{Var(Y_i)} \left(\frac{d\mu_i}{d\eta_i}\right) = 0$$

Função desvio (*deviance*) e bondade de ajuste

$$D_M = 2\log(L_S - L_M)$$

L_S é a função de verossimilhança do modelo saturado avaliada no seu ponto de máximo;
L_M é a função de verossimilhança do modelo de interesse avaliada no seu ponto de máximo.

- Valores baixo de D_M indicam um bom ajuste do modelo.
- Assintoticamente, D_M tem distribuição qui-quadrado com $\nu = N - q$ graus de liberdades. Em que N é o número de parâmetros no modelo saturado (completo) e q no modelo M.

Função desvio (*deviance*) e bondade de ajuste

Exemplo 1: O objetivo do estudo é investigar o hábito de roncar como um fator de risco para doenças cardíacas.

Hábito de roncar	Doença cardíaca		Proporção
	Sim	Não	de sim
Nunca	24	1355	0.017
Ocasional	35	603	0.055
Quase toda noite	21	192	0.099
Toda noite	30	224	0.118

Função desvio (*deviance*) e bondade de ajuste

O modelo saturado é o produto de 4 binomiais com probabilidades de sucesso diferentes. Assim $N = 4$ e

$$\log L_S = C + 24\log(0.017) + 35\log(0.055) + 21\log(0.099) + 30\log(0.118) +$$

$$+ 1355\log(1-0.017) + 603\log(1-0.055) + 192\log(1-0.099) + 224\log(1-0.118)$$

Vamos considerar o modelo sob a hipótese de homogeneidade $H_0 : p_{(1)1} = p_{(2)1} = p_{(3)1} = p_{(4)1}$. Neste caso, a log-verossimilhança é

$$C + (24 + 35 + 21 + 30)\log(\pi) + (1355 + 603 + 192 + 224)\log(1-\pi)$$

Função desvio (*deviance*) e bondade de ajuste

O ponto de máximo da verossimilhança restrita a H_0 é $\hat{\pi} = \frac{110}{2484}$.
Logo,

$$\log L_M = C + 110 \log\left(\frac{110}{2484}\right) + 2374 \log\left(\frac{2374}{2484}\right)$$

Resulta em

$$D_M = 2(\log L_S - \log L_M) = 65.89$$

O número de graus de liberdades é $\nu = 4 - 1 = 3$.
Valores altos da função desvio indicam contra o modelo.

$$P(\chi^2_3 > 65.89) = 3.23 \times 10^{-14}.$$

Comparação de modelos

- Dois modelos são ditos encaixados se M_0 é um caso particular de M_1, isto é, alguns parâmetros de M_1 são zerados para obter M_0
- Neste caso, uma medida de comparação de modelos é dada pela diferença entre as funções desvios: $D_0 - D_1$.
- Note que

$$D_0 - D_1 = 2(\log(L_S) - \log(L_{M_0})) - 2(\log(L_S) - \log(L_{M_1})) =$$

$$2 \log \left(\frac{L_{M_1}}{L_{M_0}} \right) = Q_{RV}$$

Exemplo 3: Os dados no arquivo *Evolution.dat* (Agresti, 2019) referem-se à uma pesquisa feita nos Estados Unidos com 1065 indivíduos para saber sua opinião sobre a teoria da evolução. Os indivíduos responderam a seguinte pergunta: *Human beings, as we know today, developed from earlier species of animals. True or false?* . Os pesquisadores desejam verificar se a resposta à esta questão esta associada com a ideologia politica do indivíduo. Para isto, eles foram classificados em sete níveis ideológicos; em que 1 representa extremamente conservador e 7 extremamente liberal.

Analisando as saídas do **R** temos:

```
> n <- Evo$true + Evo$false # binomial sample sizes
> fit <- glm(true/n ~ ideology, family=binomial, weights=n, data=Evo)
> summary(fit) # logistic regression fit
             Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  -1.75658     0.20500   -8.569    <2e-16
ideology      0.49422     0.05092    9.706    <2e-16 # z Wald test
---
    Null deviance: 113.20 on 6 degrees of freedom # explained in Sec. 3.4.1
Residual deviance:   3.72 on 5 degrees of freedom
Number of Fisher Scoring iterations: 3 # explained in Section 3.5.1

> confint(fit) # profile likelihood CI
               2.5 %    97.5 %
ideology     0.39617   0.59594

> library(car)
> Anova(fit) # likelihood-ratio tests for effect parameters in a GLM
         LR Chisq  Df  Pr(>Chisq)
ideology   109.48   1   < 2.2e-16 # can also get with drop1(fit, test="LRT")
> library(statmod)
> fit0 <- glm(true/n ~ 1, family=binomial, weights=n, data=Evo) # null model
> glm.scoretest(fit0, Evo$ideology)^2 # squaring a z score statistic
[1] 104.101 # score chi-squared statistic with df=1
```

- O $IC(0.95)$ de Wald para β é $(0.494 - 1.96 \times 0.051; 0.494 + 1.96 \times 0.051) = (0.394; 0.594)$.
- A função desvio associada ao modelo nulo (modelo só com o intercepto) é $D_0 = 113.2$ com 6 graus de liberdades.
- A função desvio associada ao modelo ajustado (residual) é $D_1 = 3.72$ com 5 graus de liberdades.
- A diferença entre as funções desvios é $113.2 - 3.72 = 109.48$. Valor alto indicando que o modelo 1 é melhor que o modelo 0 .
- Esse resultado coincide com a estatística do teste de Razão de Verossimilhança, obtido pelo função *Anova()* , para testar $H_0 : \beta = 0$.

Resíduos de Pearson:

$$e_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{var}(y_i)}} \quad , \quad i = 1, \ldots, n$$

Resíduos padronizados:

$$r_i = \frac{y_i - \hat{\mu}_i}{ep(y_i - \hat{\mu}_i)} \quad , \quad i = 1, \ldots, n$$

Para MLG, podemos mostrar que

$$ep(y_i - \hat{\mu}_i) = \sqrt{\hat{var}(y_i)(1 - h_i)}$$

- O denominador dos resíduos padronizados leva em consideração a variabilidade de ambos y_i e $\hat{\mu}_i$ e por isso é preferível em relação aos resíduos e_i.
- r_i tem distribuição aproximadamente normal padrão quando μ_i é grande.
- Devido a normalidade assintótica de r_i é possível estabelecer limites para os valores esperados desses resíduos.
- Devemos ter atenção para valores de $|r_i| > 2$ (pouco prováveis).

Para o caso especial de tabelas $r \times c$ com $N_{ij} \sim Poisson(\mu_{ij})$, temos que $var(N_{ij}) = \mu_{ij}$. Assim,

$$e_{ij} = \frac{N_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}}}.$$

Além disso, se o modelo ajustado é o de independência então

$$\hat{\mu}_{ij} = \frac{n_{i+} n_{+j}}{n_{++}}.$$

Resulta que a soma dos resíduos ao quadrado é a estatística qui-quadrado de Pearson

$$\sum_{i=1}^{r} \sum_{j=1}^{c} e_{ij}^2 = Q_P$$

```
> attach(Evo)
> cbind(ideology,true,false,n,true/n,fitted(fit),rstandard(fit,type="pearson"))
>       # rstandard(fit, type="pearson") requests standardized residuals
 ideology  true false    n # sample  fitted  std. res.
1        1    11    37   48   0.2292  0.2206   0.1611 # extremely conservative
2        2    46   104  150   0.3067  0.3169  -0.3515
3        3    70    72  142   0.4930  0.4319   1.6480
4        4   241   214  455   0.5297  0.5549  -1.4995
5        5    78    36  114   0.6842  0.6714   0.3249
6        6    89    24  113   0.7876  0.7701   0.5414
7        7    36     6   42   0.8571  0.8459   0.2207 # extremely liberal
```