Análise de Dados Categorizados - Aula 8

Márcia D Elia Branco

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

1. Modelos Lineares

$$Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

ϵ é uma v.a. com $E[\epsilon] = 0$.

$$\mu = E[Y \mid x] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

2. Modelos Lineares Generalizados

$$g(\mu) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

Modelos Lineares Generalizados (MLG)

- Os MLG possuem três componentes: o modelo probabilístico, o preditor linear e a função de ligação.
- O modelo probabilístico é a distribuição associada a variável resposta Y.
- O preditor linear é $\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$.
- A função de ligação é $g(\mu)$, em que μ é a valor esperado de Y.
- A função no programa **R** que ajusta MLG é

$$glm(y \sim x, family =< modeloprob. > (link =< nome >), data)$$

Casos especiais de MLG

(a) Modelo linear normal

Y é normal com média $\mu = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$.

Neste caso a função de ligação é a identidade $g(\mu) = \mu$.

(b) Modelo log linear Poisson

Y tem distribuição de Poisson com média μ e função de ligação

$$\log(\mu) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

Esta modelagem garante que a condição $\mu > 0$, sem a necessidade de adicionar alguma restrição aos parâmetros do modelo $(\alpha, \beta_1, \ldots, \beta_p)$.

(c) Modelo logístico Bernoulli (dados dicotômicos)

Y tem distribuição Bernoulli(π), então $\mu = \pi$ com $0 < \pi < 1$.

A função de ligação é

$$\log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

O logaritmo da chance também denominado *logito*.

Invertendo, temos

$$\pi = \frac{exp\{\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p\}}{1 + exp\{\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p\}}$$

Modelo de regressão logística com um preditor

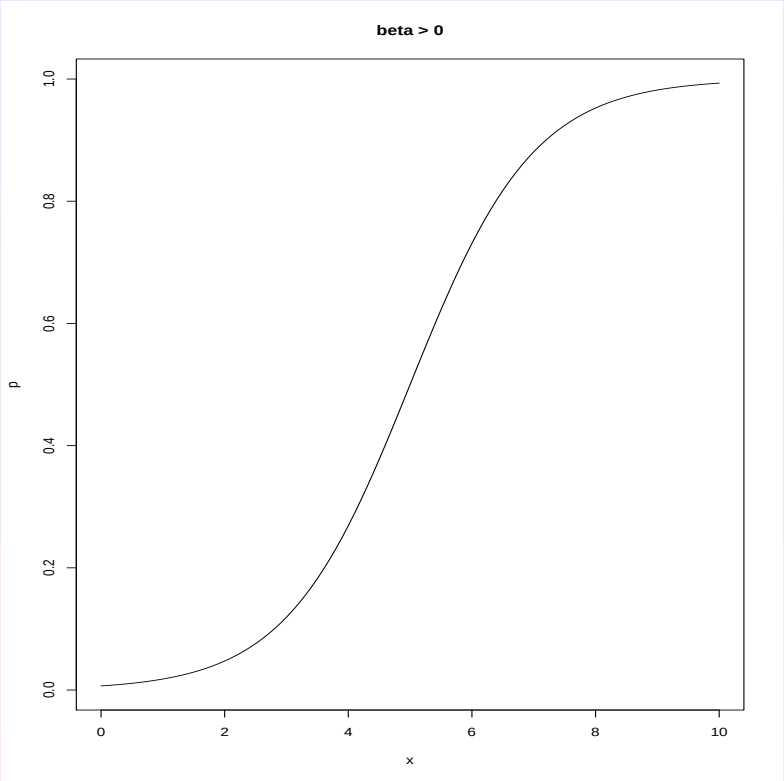
Considere $p = 1$. Então $P(Y = 1) = \pi$ é modelada por

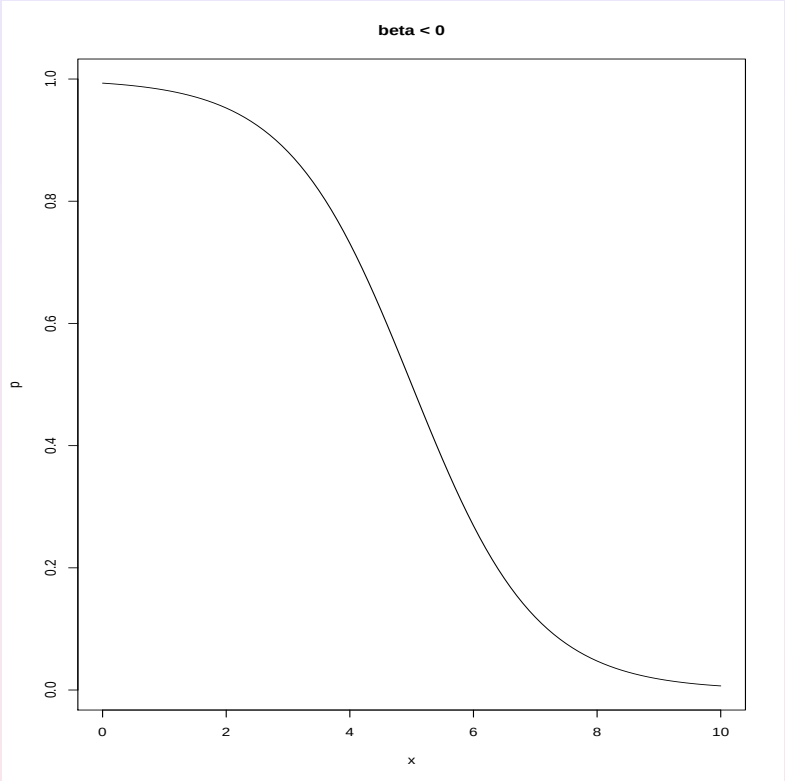
$$\pi = \frac{exp\{\alpha + \beta x\}}{1 + exp\{\alpha + \beta x\}}$$

- A função acima garante que π fica restrita ao intervalo $(0, 1)$.
- Para $\beta > 0$ a função é estritamente crescente em x. Para $\beta < 0$ a função é estritamente decrescente em x.
- O parâmetro β pode ser interpretado como: para a mudança de uma unidade na covariavel x temos, em média, o aumento (redução) de β unidades no logito .

$$logito(\pi) = \alpha + \beta x$$

Modelo de regressão logística com um preditor





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

Modelo de regressão logística com um preditor

Vamos considerar independência entre os indivíduos, $y_i = 1$ se i-ésimo indivíduo tem alguma doença cardíaca e $y_i = 0$ caso contrário.

A variável explicativa é categórica do nível ordinal. Vamos substituir as categorias por escores: $x = 0$ (nunca); $x = 2$ (ocasional); $x = 4$ (quase sempre) e $x = 5$ (sempre).

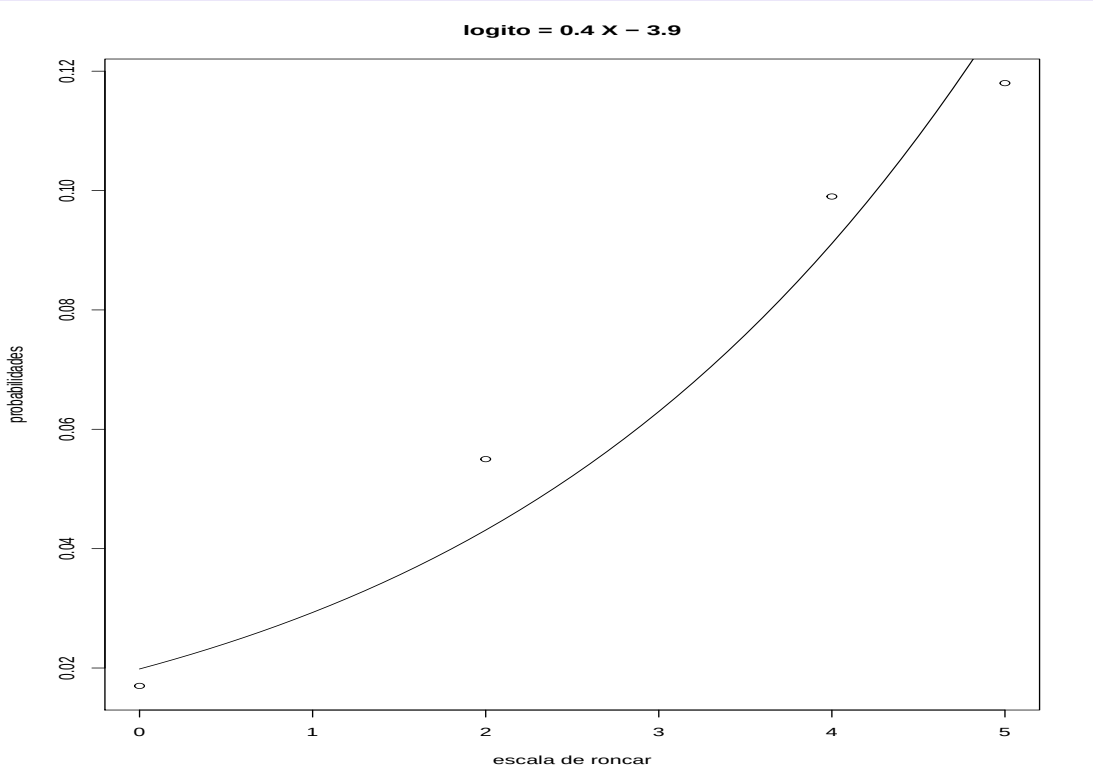
O modelo logístico ajustado foi:

$$logito(\pi) = -3.866 + 0.397x$$

Equivalentemente

$$\hat{\pi} = \frac{e^{-3.866+0.397x}}{1 + e^{-3.866+0.397x}}$$

Modelo de regressão logística com um preditor



Modelo de regressão logística com um preditor

- O valor estimado $\hat{\beta} > 0$ indica que a probabilidade de doença cardíaca aumenta com o aumento do hábito de roncar.
- O aumento médio no logito é aproximadamente $\hat{\beta} = 0.4$ para o aumento de uma unidade na escala de roncar.
- Melhor interpretar em termos de chance. Assim, a chance de doença cardíaca aumenta, em média, $e^{0.4} = 1.49$ com o aumento de uma unidade na escala de roncar.
- As estimativas $\hat{\beta}$ e $\hat{\alpha}$ são obtidas via método de máxima verossimilhança e uso de algoritmos numéricos.

MLG estimação dos parâmetros

O modelo probabilístico de um MLG deve pertencer a família exponencial.

A família exponencial é caracterizada pela seguinte função de probabilidade (densidade).

$$f(y_i;\theta_i,\phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right\} \quad , i = 1,\ldots,n$$

θ_i é denominado parâmetro natural e
ϕ é um parâmetro de dispersão, que não depende de i.

A função de ligação canônica é dada por $g(\mu_i) = \theta_i$.

Caso especial: $y_i \sim Bernoulli(\pi_i)$ independentes.
Neste caso $\phi = 1$ e

$$f(y_i;\theta_i) = \exp\left\{y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + \log(1-\pi_i)\right\} \quad , i = 1, \ldots, n$$

$\theta_i = \log\left(\frac{\pi_i}{1-\pi_i}\right)$ é o logito ou logaritmo da chance.
A log-verossimilhança é dada por

$$\log(L(\theta)) = \sum_{i=1}^{n} y_i\theta_i - \log(1+e^{\theta_i})$$

Derivando em θ_i temos

$$y_i + \frac{db(\theta_i)}{d\theta_i}$$

em que $\frac{db(\theta_i)}{d\theta_i}$ é a primeira derivada de $b(\theta_i) = -\log(1 + e^{\theta_i})$.

Resulta que a derivada da log-verossimilhança é dada por

$$\frac{d \log L_i}{d\theta_i} = y_i - \frac{e^{\theta_i}}{1 + e^{\theta_i}} \tag{1}$$

Considere agora um preditor linear da forma $\alpha + \beta x_i$ e

$$\theta_i = \alpha + \beta x_i$$

Para obter os emv de α e β devemos derivar a log-verossimilhança em ambos. Vamos obter primeiro as equações de estimação para β. Assim, derivando em β

$$\frac{d \log L(\alpha, \beta)}{d\beta} = \sum_{i=1}^{n} \frac{d \log L_i}{d\beta} \tag{2}$$

Pela regra da cadeia

$$\frac{d \log L_i}{d\beta} = \left(\frac{d \log L_i}{d\theta_i}\right)\left(\frac{d\theta_i}{d\beta}\right)$$

Mas $\frac{d\theta_i}{d\beta} = x_i$. Portanto, de (1) temos que

$$\frac{d \log L_i}{d\beta} = x_i y_i - x_i \frac{e^{\theta_i}}{1 + e^{\theta_i}} = x_i y_i - x_i \left(\frac{e^{\alpha+\beta x_i}}{1 + e^{\alpha+\beta x_i}} \right)$$

Voltando a (2), temos

$$\frac{d \log L(\alpha, \beta)}{d\beta} = \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \left(\frac{e^{\alpha+\beta x_i}}{1 + e^{\alpha+\beta x_i}} \right) \qquad (3)$$

Analogamente, derivando em α obtemos

$$\frac{d \log L(\alpha, \beta)}{d\alpha} = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \left(\frac{e^{\alpha+\beta x_i}}{1 + e^{\alpha+\beta x_i}} \right) \qquad (4)$$

MLG estimação dos parâmetros: modelo Bernoulli

- Para obtenção dos emv deve-se igualar a zero (3) e (4) e resolver o sistema de equações.
- A solução é obtida pelo uso do algoritmo de Newton-Raphson ou Escore de Fisher .
- A inferência clássica é baseada na distribuição assintótica normal dos estimadores de máxima verossimilhança.
- A variância assintótica é obtida pela inversa da matriz de informação de Fisher.

MLG estimação dos parâmetros: modelo Bernoulli

- Para obtenção da matriz de variância assintótica dos emv, precisamos da matriz de informação de Fisher:

$$I_F(\alpha,\beta) = -E\left[\begin{pmatrix} \frac{d^2 \log L}{d\alpha d\alpha} & \frac{d^2 \log L}{d\alpha d\beta} \\ \frac{d^2 \log L}{d\alpha d\beta} & \frac{d^2 \log L}{d\beta d\beta} \end{pmatrix}\right]$$

Obtendo as segundas derivadas.

$$\frac{d^2 \log L}{d\alpha d\alpha} = -\sum_{i=1}^{n} \frac{d\left(\frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}\right)}{d\alpha} = -\sum_{i=1}^{n} \left(\frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}\right)^2$$

$$\frac{d^2 \log L}{d\alpha d\beta} = -\sum_{i=1}^{n} \frac{d\left(\frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}\right)}{d\beta} = -\sum_{i=1}^{n} x_i \left(\frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}\right)^2$$

$$\frac{d^2 \log L}{d\beta d\beta} = -\sum_{i=1}^{n} \frac{d\left(x_i \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}\right)}{d\beta} = -\sum_{i=1}^{n} x_i^2 \left(\frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}\right)^2$$

MLG estimação dos parâmetros: modelo Bernoulli

- Nenhuma das segundas derivadas depende de Y_i. Portanto, o valor esperado é a própria derivada.
- Para construção da $I_F(\alpha, \beta)$ basta mudar o sinal das expressões das segundas derivadas.
- A matriz de variância assintótica é obtida pela inversa da matriz de informação: $[I_F(\alpha, \beta)]^{-1}$.
- Note que essa matriz depende dos parâmetros desconhecidos. Substituindo-se pelas estimativas de máxima verossimilhança, temos a aproximação usado nos intervalos e testes de Wald.

Testes de hipóteses

Na maioria dos casos a hipótese de interesse é $H_0 : \beta = 0$ versus $H_a : \beta \neq 0$. Sob H_0, para amostras grandes,

$$Z = \frac{\hat{\beta}}{ep(\hat{\beta})}$$

tem distribuição aproximadamente normal padrão.
Assim, Z^2 tem distribuição qui-quadrado com $\nu = 1$ g.l..

Inferência em MLG com $p = 1$

Os testes de Wald e Escore se diferenciam pela maneira de estimar

$$ep(\hat{\beta}) = \sqrt{Var[\hat{\beta}]}$$

- A variância assintótica é obtida pela inversa da matriz de informação de Fisher.
- Esta por sua vez, depende do negativo da segunda derivada da log-verossimilhanca.
- A matriz de informação pode depender do parâmetro desconhecido β. Neste caso, ele será substituído pelo emv (Wald) ou zero (Escore).

Inferência em MLG com $p = 1$

- O teste da razão de verossimilhança é baseado em $2[\log(L_1) - \log(L_0)]$, que no nosso caso tem distribuição aproximada qui-quadrado com $\nu = 1$ g.l.
- As estimativas dos erros padrões e da estatística da RV são obtidas por aproximações numericas.
- O erro padrão calculado pela função *glm* do **R** é o de Wald.
- A estatística de verossimilhança pode ser obtida usando-se *library (car)* e a função *Anova*
- Para o teste escore pode-se usar *library(statmod)* e a função *glm.scoretest* .

Exemplo 1: regressão logística

Para os dados do exemplo1, vamos testar

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

Para isso vamos considerar a estatística do teste de Wald, dada por

$$z^2 = \left(\frac{0.3974}{0.05}\right)^2 = 63.15$$

A região critica do teste baseada na distribuição qui-quadrado com 1 g.l. e $\alpha = 0.05$ é $[3.84, \infty)$. Rejeita-se $H_0 (Valor - P < 10^{-15})$.

Conclusão: Existe efeito do hábito de roncar na probabilidade do indivíduo ter doença cardíaca.

Saída do *glm* para exemplo 1

```
> Heart <- read.table("http://www.stat.ufl.edu/~aa/cat/data/Heart.dat",
+                     header=TRUE)
> Heart # Heart data file at text website, in contingency table form
           snoring yes   no
1            never  24 1355
2        occasional  35  603
3 nearly_every_night  21  192
4        every_night  30  224
> # the following code fits logistic regression model to the data file
> library(dplyr) # to recode explanatory variable
> Heart$x <- recode(Heart$snoring, never = 0, occasional = 2,
+                   nearly_every_night = 4, every_night = 5)
> n <- Heart$yes + Heart$no # binomial sample sizes are the row totals
> fit <- glm(yes/n ~ x, family=binomial(link=logit), weights=n, data=Heart)
> # canonical link for binomial is logit, so "(link=logit)" not necessary
> # "weights" indicates sample proportion yes/n is based on n observations
> summary(fit)
             Estimate   Std. Error
(Intercept)  -3.86625      0.16621
x             0.39734      0.05001 # logistic ML estimate of beta is 0.397
> fitted(fit) # fitted values (probability estimates) at 4 levels of snoring
      1       2       3       4
0.02051 0.04430 0.09305 0.13244
```

Exemplo 1: regressão logística

Os valores estimados para as probabilidades π_i via modelo são obtidos por

$$\hat{\pi} = \frac{exp\{-3.866 + 0.397x\}}{1 + exp\{-3.866 + 0.397x\}}$$

Como $x = (0, 2, 4, 5)$, temos

$$\hat{\pi}_1 = 0.021 \ , \ \hat{\pi}_2 = 0.0443 \ , \ \hat{\pi}_3 = 0.093 \ , \ \hat{\pi}_4 = 0.132$$

O Intervalo de 95 % de confiança de Wald para β é obtido por:

$$(\hat{\beta} - 1.96ep(\hat{\beta}); \hat{\beta} - 1.96ep(\hat{\beta})) = (0.299; 0.495)$$

O intervalo de 95 % de confiança de Wald para α é obtido por:

$$(\hat{\alpha} - 1.96ep(\hat{\alpha}); \hat{\alpha} - 1.96ep(\hat{\alpha})) = (3.54; 4.19)$$

Regressão logística em tabelas 2×2

Considere $x = (0, 1)$ e

$$logito(\pi(x)) = \alpha + \beta x$$

Então :

$$logito(\pi(0)) = \alpha \quad \text{e} \quad logito(\pi(1)) = \alpha + \beta$$

Portanto

$$\beta = logito(\pi(1)) - logito(\pi(0)).$$

Mas

$$logito(\pi(x)) = \log\left(\frac{\pi(x)}{1 - \pi(x)}\right)$$

Portanto

$$\beta = \log\left(\frac{\pi(1)[1-\pi(0)]}{[1-\pi(1)]\pi(0)}\right)$$

Neste caso o parâmetro β representa o logaritmo da razão de chances (OR) . E

$$OR = e^{\beta}$$