Análise de Dados Categorizados - Aula 10

Márcia D Elia Branco

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

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.

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:

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

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

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

Tabelas $2 \times 2 \times 2$

- Y é a variável resposta com apenas dois valores (0-1).
- As variáveis explicativas são também dicotomizadas. Assim: $x_i = 1$ se ocorre o evento de interesse e $x_i = 0$ caso contrário, para $i = 1, 2$.
- O modelo de regressão logística é dada por

$$logito(\pi(x_1, x_2)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

- Interpretação dos parâmetros: β_0 é o valor do logito na casela de referência $(x_1 = 0, x_2 = 0)$; β_1 é o incremento no logito para mudança na variável x_1 de 0 para 1, mantido fixo x_2; e β_2 é o incremento no logito para mudança na variável x_2 de 0 para 1, mantido fixo x_1 .

Exemplo 1: Os dados na tabela a seguir mostram resultado de um estudo sobre doença coronária. Os pacientes foram submetidos a um eletrocardiograma (ECG) e classificados conforme o resultado da medição (0 se < 0.1 e 1 se ≥ 0.1). Além disso a variável sexo também foi controlada (0 se feminino e 1 se masculino).

Sexo	ECG	Doença		Totais
		Presente	Ausente	
Feminino	< 0.1	4	11	15
	≥ 0.1	8	10	18
Masculino	< 0.1	9	9	18
	≥ 0.1	21	6	27

Temos que $Y = 1$ se a doença é presente e $Y = 0$ se esta ausente.

A casela de referência é dada por $x_1 = 0$ e $x_2 = 0$, representada por Feminino e ECG < 0.1 .

As estimativas de máxima verossimilhança com seus respectivos erros são:

$$\hat{\beta}_0 = -1.17(0.485) \quad \hat{\beta}_1 = 1.28(0.498) \quad \hat{\beta}_2 = 1.05(0.498)$$

O modelo ajustado é

$$logito(\pi(x_1, x_2)) = -1.17 + 1.28x_1 + 1.05x_2$$

Regressão logística em tabelas de contingência

- A chance de doença é obtida por $e^{\beta_0+\beta_1 x_1+\beta_2 x_2}$.
- Para $x_2 = 0$ a razão de chances entre $x_1 = 1$ e $x_1 = 0$ fica expressa por $\frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}$.
- Para $x_2 = 1$ a razão de chances entre $x_1 = 1$ e $x_1 = 0$ fica expressa por $\frac{e^{\beta_0+\beta_1+\beta_2}}{e^{\beta_0+\beta_2}} = e^{\beta_1}$.
- Assim, o modelo tem a restrição de que as razões de chances não dependem de x_2 .
- De modo similar notamos que a razão de chances entre $x_2 = 1$ e $x_2 = 0$ fica expressão por e^{β_2} para ambos valores de x_1

Para o nosso exemplo, as estimativas das razões de chances (OR) via modelo de regressão logística são:

Entre Masculino e Feminino: $e^{1.28} = 3.60$.

Entre ECG alto e ECG baixo: $e^{1.05} = 2.87$.

De um modo geral os resultados indicam que indivíduos do sexo masculino com ECG alto são mais propensos a apresentar doença coronária.

O intervalo de confiança (0.95) de Wald para β_1 é

$$(1.28 - 1.96 \times 0.498; 1.28 - 1.96 \times 0.498) = (0.304; 2.256)$$

O intervalo de Wald de confiança 0.95 para $OR_{M/F}$ é dado por

$$(e^{0.304}; e^{2.256}) = (1.35; 9.54)$$

O intervalo de Wald de confiança 0.95 para $OR_{ECGA/ECGB}$ é dado por

$$(e^{1.05-1.96\times0.498}; e^{1.05+1.96\times0.498}) = (1.08; 7.58)$$

Como incluir a interação no modelo?

$$logito(\pi(x_1, x_2)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

Assim

(x_1, x_2)	Chance
(0,0)	$\exp(\beta_0)$
(0,1)	$\exp(\beta_0 + \beta_2)$
(1,0)	$\exp(\beta_0 + \beta_1)$
(1,1)	$\exp(\beta_0 + \beta_1 + \beta_2 + \beta_3)$

Como incluir a interação no modelo?

$$logito(\pi(x_1, x_2)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

Assim

(x_1, x_2)	Chance
(0,0)	$\exp(\beta_0)$
(0,1)	$\exp(\beta_0 + \beta_2)$
(1,0)	$\exp(\beta_0 + \beta_1)$
(1,1)	$\exp(\beta_0 + \beta_1 + \beta_2 + \beta_3)$

$$\hat{OR}_{x_2|x_1=0} = \exp(\beta_2) \qquad \hat{OR}_{x_2|x_1=1} = \exp(\beta_2 + \beta_3)$$

As razões de chances agora dependem da segunda variável considerada.

$$\hat{OR}_{x_1|x_2=0} = \exp(\beta_1) \qquad \hat{OR}_{x_1|x_2=1} = \exp(\beta_1 + \beta_3)$$

Estratégia de análise:

1. Testar $\beta_3 = 0$. Se a hipótese não for rejeitada, ajustar o modelo apenas com os efeitos principais.
2. Após o segundo ajuste, testar $\beta_1 = 0$ e $\beta_2 = 0$. Manter no modelo apenas as variáveis associadas ao teste que rejeita a hipótese nula.

Para os dados do exemplo 1, temos o seguinte resultado obtido pela função *glm()* :

Modelo	g.l.	*Deviance*	Dif *Dev*	Dif g.l.	Valor-P
Nulo	3	11.98			
X_1	2	4.86	7.12	1	0.0076
$X_2 \mid X_1$	1	0.21	4.65	1	0.0311
$X_1 * X_2 \mid X_1, X_2$	0	0.000	0.21	1	0.6436

Notamos que a diferença entre as funções desvios do modelo sem interação e o modelo com interação é 0.21. Este valor é o mesmo da estatística do teste de razão de verossimilhança para testar $H_0 : \beta_3 = 0$. Como o Valor-P=0.6436, não rejeita-se H_0. Podemos excluir a interação do modelo.

A diferença das funções desvios entre o modelo sem a variável *ECG* e com a variável *ECG* é 4.65 com um Valor-P = 0.0311. Indicando a rejeição de $H_0 : \beta_2 = 0$. Mantemos a variável *ECG* no modelo.

A diferença das funções desvios entre o modelo sem a variável *sexo* e com a variável *sexo* é 7.12 com um Valor-P = 0.0076. Indicando a rejeição de $H_0 : \beta_1 = 0$. Mantemos a variável *sexo* no modelo.

Os testes considerados na nossa análise são da RV. Poderia ter sido considerado também os testes de Wald cujos resultados são obtidos diretamente na saída da função *summary(ajuste)* .

Regressão binária versus Regressão Binomial

Regressão binária : Considera $Z_i \sim Bernoulli(\pi_i) \quad i = 1, \ldots, n$. Denominada dados não agrupados.

- Número de parâmetros associado ao modelo saturado é n.
- As estimativas pontuais de mv das média são $\hat{\mu}_i = z_i$. Somente valores 0 ou 1.
- Considerando o modelo de regressão (MLG)

$$g(\pi(x_i)) = \beta x_i,$$

as estimativas para as probabilidades via modelo são $\hat{\pi}(x_i) = g^{-1}(\hat{\beta} x_i)$, em que $\hat{\beta}$ é o emv de β.

- Note que as estimativas via modelos podem ser qualquer valor no intervalo (0,1) .

Regressão binária versus Regressão Binomial

- A função desvio associada ao modelo com p preditores linear tem $\nu = n - (p+1)$ graus de liberdades. Precisamos garantir que $n > p+1$.
- A função desvio é dada por

$$D_M = 2\sum_{i=1}^{n} z_i[\log z_i - \log \hat{\pi}(x_i)] + (1-z_i)[\log(1-z_i) - \log(1-\hat{\pi}(x_i))]$$

Regressão binária versus Regressão Binomial

Regressão Binomial : $Y_j \sim Binomial(n_j, \pi_j)$ $j = 1, \ldots, k$.
Denomindada dados agrupados.

- Número de parâmetros associado ao modelo saturado é $k < n = \sum_{j=1}^{k} n_j$.
- As estimativas pontuais das médias são $\hat{\mu}_j = y_j$. Valores inteiros entre 0 e n_j .
- As estimativas pontuais de mv para os π_j são $\frac{y_j}{n_j}$.
- Considerando o modelo de regressão (MLG)

$$g(\pi(x_j)) = \beta x_j \quad , \quad j = 1, \ldots k.$$

as estimativas para as probabilidades via modelo são $\hat{\pi}(x_j) = g^{-1}(\hat{\beta} x_j)$, em que $\hat{\beta}$ é o emv de β.

Regressão binária versus Regressão Binomial

- A função desvio associada ao modelo com p preditores linear tem $\nu = k - (p+1)$ graus de liberdades. Precisamos garantir que $k > p+1$.
- A função desvio é dada por

$$2\sum_{j=1}^{k} y_j \left[\log\left(\frac{y_j}{n_j - y_j}\right) - \log\hat{\pi}(x_j)\right] + (n_j - y_j)\left[\log\left(n_j - \frac{y_j}{n_j}\right) - \log\left(1 - \hat{\pi}(x_j)\right)\right]$$

Regressão binária versus Regressão Binomial

- As funções desvios para os modelo binário e binomial são diferentes. (exercício!)
- Vamos obter valores diferentes de função desvio dependendo de como ingressamos os dados, agrupados ou não-agrupados.
- No entanto se tivermos dois modelo encaixados M_0 e M_1, a diferença das funções desvios para os dois casos será a mesma.
- Lembre que $D_{M_0} - D_{M_1} = 2(\log L_{M_1} - \log L_{M_0})$, não depende da verossimilhança no modelo saturado.
- Para x_j fixado, temos que $y_j = \sum_{i=1}^{n_j} z_{ji}$. Em que z_{ij} são respostas 0-1 (Bernoulli).

Então, para o modelo binomial

$$\log L_{M_1} = \sum_{j=1}^{k} y_j \log\left(\frac{\hat{\pi}(x_j)}{1-\hat{\pi}(x_j)}\right) + \sum_{j=1}^{k} n_j \log(1-\hat{\pi}(x_j)) =$$

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j} z_{ij} \log\left(\frac{\hat{\pi}(x_j)}{1-\hat{\pi}(x_j)}\right) + \sum_{j=1}^{k}\sum_{i=1}^{n_j} \log(1-\hat{\pi}(x_j))$$

Fazendo um pequeno ajuste de notação, a expressão acima é igual a $\log L_{M_1}$ obtida usando o modelo Bernoulli.

Equivalentemente para $\log L_{M_0}$.

Exercícios

1) Considere o seguinte conjunto de dados artificiais: (0,4,1), (1,4,2) e (2,4,4). Em que o primeiro valor corresponde à variável explicativa X, o segundo é o número de ensaios (tamanho da amostra binomial) e o terceiro valor é o número de sucessos. Usando o **R**, entre com os dados de duas maneiras: (i) não agrupados, considerando as 12 observações separadamente e (ii) agrupados, considerando as 3 binomiais. Para ambos ajuste um modelo de regressão logística.

(a) Apresente a saída do ajuste obtida pela função *glm*. Compare os valores obtidos das funções desvio dos modelos nulos e completos em ambos os casos. Esses valores dependeram da entrada dos dados (i) ou (ii) ?

Exercícios

Saída: dados não agrupados

```
Call:
glm(formula = y ~ x, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4216  -0.6339   0.3752   0.5193   1.8459

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.503      1.181  -1.272   0.2033
x              2.060      1.130   1.823   0.0682 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 16.301  on 11  degrees of freedom
Residual deviance: 11.028  on 10  degrees of freedom
AIC: 15.028
```

Saída: dados agrupados

```
Call:
glm(formula = cbind(rsim, rnao) ~ x, family = binomial)

Deviance Residuals:
      1       2       3
 0.3377  -0.5543   0.7504

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.503      1.181  -1.272   0.2034
x              2.060      1.130   1.823   0.0683 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6.2568  on 2  degrees of freedom
Residual deviance: 0.9844  on 1  degrees of freedom
AIC: 8.6722
```

Resposta: os valores em amarelo representam as funções desvios do modelo nulo e do modelo de regressão logistica. Nota-se que os valores são distintos e portanto, dependeram da entrada dos dados.

(b) Obtenha agora a diferença das funções desvio (Modelo Nulo - Modelo completo) nos dois tipos de entrada de dados. O que pode ser notado? Qual a relação destas medidas com o Teste da Razão de Verossimilhança?

Para dados não agrupados: D0-D1= 5.273
Para dados agrupados: D0-D1=5.2724

Obs:As medidas devem ser iguais. A pequena diferença que aparece é devido a arredondamento de valores.

As diferenças desses desvios é o valor da estatística do teste de razão de verossimilhança associado as hipóteses $H_0 : \beta = 0$ versus $H_a : \beta \neq 0$

Exercícios

2) Para avaliar a toxidade de um inseticida, um bioensaio foi conduzido. Doses crescentes do inseticida foram aplicadas a grupos de insetos, registrando-se o número de mortes em cada grupo. Foi ajustado um modelo de regressão logística e os resultados são apresentados a seguir.

```
> dose=c(0,2.6,3.8,5.1,7.7,10.2)
> nsim=c(0,6,16,24,42,44)
> nnao=c(49,44,32,22,7,6)
> ajuste=glm(cbind(nsim,nnao)~dose, family=binomial)
> summary(ajuste)

Call:
glm(formula = cbind(nsim, nnao) ~ dose, family =
binomial)

Deviance Residuals:
      1       2       3       4       5       6
-1.9540  -0.8157   0.7507   0.7679   0.9145  -1.9456

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.22566    0.36992  -8.720   <2e-16 ***
dose         0.60513    0.06781   8.923   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 163.745  on 5  degrees of freedom
Residual deviance:  10.258  on 4  degrees of freedom
AIC: 33.479

Number of Fisher Scoring iterations: 5
```

Exercícios

(a) Escreva o modelo ajustado. Interprete os valores-P associados a cada coeficiente e especifique qual o teste de hipótese que esta sendo considerado.

(b) Obtenha o valor predito de probabilidade de morte associado a $dose = 3.8$.

(c) Determine as doses letais 50 % e 90 %, isto é, os valores estimados de doses associadas as probabilidades de morte 0.5 e 0.9 .

Solução:

(a) Modelo ajustado: $logito(p(x)) = -3.226 + 0.605x$ em que

$$logito(p(x)) = \log\left(\frac{p(x)}{1-p(x)}\right) \quad \text{e} \quad x \text{ é a dose de inseticida.}$$

As hipóteses testadas são:
- $H_0 : \beta_0 = 0$, resultando num valor-P $< 10^{-16}$. Rejeita-se H_0
- $H_0 : \beta_1 = 0$, resultando num valor-P $< 10^{-16}$. Rejeita-se H_0

Ambos os coeficientes são significativos.

(b)

$$\hat{p}(3.8) = \frac{e^{-3.22566+0.60513\times 3.8}}{1+e^{-3.22566+0.60513\times 3.8}} = 0.2837$$

(c) Dose letal 50 % : é o valor x_{50} associado a $p = 0.5$. Temos que

$$\log\left(\frac{0.5}{0.5}\right) = 0$$

Portanto,

$$0 = -3.22566 + 0.60513 x_{50} \quad \Leftrightarrow \quad x_{50} = \frac{3.22566}{0.60513} = 5.33$$

Dose letal 90 % : é o valor x_{90} associado a $p = 0.9$. Temos que

$$\log\left(\frac{0.9}{0.1}\right) = 2.197$$

Portanto,

$$2.197 = -3.22566 + 0.60513 x_{90} \quad \Leftrightarrow$$

$$x_{90} = \frac{2.197 + 3.22566}{0.60513} = 8.96$$

Exercícios

3) Um estudo sobre bronquite (Sim ou Não) teve o objetivo de avaliar a associação entre esta doença e as variáveis $X1$ = fumo (1 se fuma e 0 se não fuma) e $X2$= idade (1 se maior ou igual a 40 anos e 0 se menos de 40 anos). Com base na análise da saída do ajuste apresentanda abaixo, responda as questões.

```
Call:
glm(formula = cbind(nsim, nnao) ~ x1 + x2 + x1 * x2,
family = binomial)

Deviance Residuals:
      1          2          3          4          5          6
7          8
 0.4928     0.4165    -0.5398    -0.4366     1.8831     1.7773
-2.1120    -1.9902

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.75199    0.14931  -5.036 4.75e-07 ***
x1           0.40683    0.18797   2.164  0.03044 *
x2           0.09353    0.19885   0.470  0.63811
x1:x2        0.73792    0.26247   2.811  0.00493 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 72.798  on 7  degrees of freedom
Residual deviance: 16.025  on 4  degrees of freedom
AIC: 65.696

Number of Fisher Scoring iterations: 3
```

Exercícios

(a) Escreva o modelo ajustado. Qual é o valor predito de probabilidade de doença para um indivíduo fumante e idade maior ou igual a 40 anos?

(b) Calcule e interprete as razões de chances de doença entre indivíduos fumante e não fumantes.

(c) Explique como foram obtidos os valores de graus de liberdades associados as funções desvio dos modelos nulo e ajustado (residual).

Solução:

(a) Modelo ajustado:
$logito(p(x)) = -0.75 + 0.407x_1 + 0.093x_2 + 0.738x_1x_2$

Valor predito para $x_1 = 1$ e $x_2 = 1$

$$\hat{p}(x) = \frac{e^{-0.75+0.407+0.093+0.738}}{1 + e^{-0.75+0.407+0.093+0.738}} = 0.619$$

(b) A razão de chances é dada por

$$\hat{OR}_{F/NF} = \frac{e^{-0.75+0.407+0.093x_2+0.738x_2}}{e^{-0.75+0.738x_2}}$$

Temos que considerar dois casos:

Se $x_2 = 0$ (menos que 40 anos), então $\hat{OR}_{F/NF} = e^{0.407} = 1.5$

Se $x_2 = 1$ (40 anos ou mais), então $\hat{OR}_{F/NF} = e^{0.407+0.738} = 3.1$

Conclusão: Para indivíduos com idade inferior a 40 anos, a chance de bronquite para fumantes é 1.5 vezes a de não fumantes. Por outro lado, para indivíduos com 40 anos ou mais, essa chance é multiplicada por 3.1 quando o indivíduo é fumante, relativamente ao não fumante.

(c) Para obter o número de graus de liberdades, primeiro temos que saber o número de parâmetros no modelo saturado.

Temos $2 \times 2 = 4$ categorias de respostas (duas variáveis explicativas). Portanto, $N = 4$

Para o modelo nulo deveriamos ter $\nu = 4 - 1 = 3$ e para o modelo ajustado $\nu = 4 - 4 = 0$

Tem algo estranho nos graus de liberdades da saída do **R**. O que pode ser?

Considera 8 categorias, pois tem 8 valores de resíduos.

Os dados incluídos consideraram mais uma variável explicativa (dicotômica) que não foi especificada no enunciado do exercício.