Análise de Dados Categorizados - Aula 11

Márcia D Elia Branco

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

Exemplo 1: O objetivo é avaliar a associação de bronquite com as variáveis X_1 : fumo (1 se fuma e 0 se não fuma), X_2 : status socioeconômico (1 se baixo e 0 se alto) e X_3 : idade (0 se abaixo de 40 anos e 1 se maior ou igual a 40).

Fumo	Socio	Idade	BRC		Totais
			Sim	Não	
0	1	0	38	73	111
0	1	1	48	86	134
0	0	0	28	67	95
0	0	1	40	84	124
1	1	0	84	89	173
1	1	1	102	46	148
1	0	0	47	96	143
1	0	1	59	53	112

Regressão Logística em Tabelas de Contingência

- Identificar a variável resposta e as variáveis explicativas.
- Qual o modelo probabilístico?
- Quantos graus de liberdades tem o modelo saturado?
- Como escrever o modelo de regressão logística mais completo?

$$logito(p(x)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 +$$

$$\beta_4 x_1 * x_2 + \beta_5 x_1 * x_3 + \beta_6 x_2 * x_3 + \beta_7 x_1 * x_2 * x_3$$

- Analisar as saídas do **R** para escolha do modelo ajustado.

Exemplo1: modelo com interação tripla

```
 Call:
glm(formula = cbind(nsim, nnao) ~ x1 + x2 + x3 + x1 * x2
+ x1 *
    x3 + x2 * x3 + x1 * x2 * x3, family = binomial)

Deviance Residuals:
[1]  0  0  0  0  0  0  0  0

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.87249    0.22503  -3.877 0.000106 ***
x1           0.15829    0.28694   0.552 0.581191
x2           0.21961    0.30109   0.729 0.465755
x3           0.13055    0.29588   0.441 0.659046
x1:x2        0.43677    0.38143   1.145 0.252177
x1:x3        0.69090    0.39377   1.755 0.079334 .
x2:x3       -0.06082    0.40003  -0.152 0.879148
x1:x2:x3     0.09353    0.53124   0.176 0.860249
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 7.2798e+01  on 7  degrees of freedom
Residual deviance: 6.2172e-14  on 0  degrees of freedom
AIC: 57.672

Number of Fisher Scoring iterations: 3
```

Exemplo 1: modelo com interações duplas

```
Call:
glm(formula = cbind(nsim, nnao) ~ x1 + x2 + x3 + x1 * x2
+ x1 *
    x3 + x2 * x3, family = binomial)

Deviance Residuals:
       1         2         3         4         5
6         7         8
 0.06640  -0.05966  -0.07442   0.06377  -0.05042
0.05879   0.05908  -0.06275

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.855760   0.203296  -4.209 2.56e-05 ***
x1           0.131035   0.241266   0.543  0.58705
x2           0.189595   0.247914   0.765  0.44441
x3           0.101563   0.245579   0.414  0.67919
x1:x2        0.484980   0.265445   1.827  0.06769 .
x1:x3        0.742284   0.264270   2.809  0.00497 **
x2:x3       -0.007801   0.263190  -0.030  0.97636
---
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:  0.031  on 1  degrees of freedom
AIC: 55.703

Number of Fisher Scoring iterations: 3
```

Exemplo 1: modelo com 2 interações duplas

```
glm(formula = cbind(nsim, nnao) ~ x1 + x2 + x3 + x1 * x2
+ x1 *x3, family = binomial)

Deviance Residuals:
        1         2         3         4         5
6         7         8
 0.07619  -0.06843  -0.08535   0.07317  -0.04071
0.04748   0.04769  -0.05066

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.8533     0.1856  -4.597 4.29e-06 ***
x1            0.1306     0.2408   0.542  0.58751
x2            0.1852     0.1982   0.934  0.35014
x3            0.0973     0.1991   0.489  0.62501
x1:x2         0.4859     0.2637   1.843  0.06536 .
x1:x3         0.7422     0.2643   2.809  0.00497 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 72.798230  on 7  degrees of freedom
Residual deviance:  0.031879  on 2  degrees of freedom
AIC: 53.704

Number of Fisher Scoring iterations: 3
```

Exemplo 1: modelo com 1 interações dupla

```
Call:
glm(formula = cbind(nsim, nnao) ~ x1 + x2 + x3 + x1 * x3,
family = binomial)

Deviance Residuals:
      1       2       3       4       5       6
7       8
-0.5238  -0.7713   0.6073   0.8498   0.5457   0.5955
-0.6301  -0.6446

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.0105     0.1683  -6.005 1.91e-09 ***
x1            0.4078     0.1891   2.156  0.03105 *
x2            0.4618     0.1307   3.534  0.00041 ***
x3            0.1036     0.2000   0.518  0.60447
x1:x3         0.7285     0.2640   2.759  0.00580 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 72.7982  on 7  degrees of freedom
Residual deviance:  3.4253  on 3  degrees of freedom
AIC: 55.097

Number of Fisher Scoring iterations: 3
```

Exemplo 1: modelo escolhido

Considerando as *deviance* residuais e o critério AIC para comparação de modelos [$AIC = -2 \log L(\theta) + 2q$], o modelo escolhido é aquele que mantem as interações duplas $Fumo * Socio$ e $Fumo * Idade$. Dado por:

$$logito[\hat{p}(x)] = -0.853 + 0.13x_1 + 0.18x_2 + 0.10x_3 + 0.49x_1 * x_2 + 0.74x_1 * x_3$$

- Quanto vale a estimativa da razão de chances entre Fumantes ($x_1 = 1$) e não fumantes ($x_1 = 0$) ?
- É unica? Depende das outras covariáveis?

Exemplo 1: modelo escolhido

Claramente o valor da razão de chances irá depender das outras covariáveis. Assim

(x_2, x_3)	$\hat{OR}_{Fumo(1/0)}$
(0,0)	$\exp\{0.13\} = 1.14$
(1,0)	$\exp\{0.13 + 0.49\} = 1.86$
(0,1)	$\exp\{0.13 + 0.74\} = 2.39$
(1,1)	$\exp\{0.13 + 0.49 + 0.74\} = 3.90$

Para indivíduos de classe socioeconômica baixa e idade alta, a chance de bronquite para um fumante é 3.9 vezes em relação aos não fumantes.
Para indivíduos de classe socioeconômica alta e idade baixa, a chance de bronquite para fumante é 1.14 vezes em relação aos não fumantes.

Regressão Logística: EMV infinito e esparsidade

- Uma tabela é dita esparsa quando possui uma grande quantidade de frequências iguais a zero.
- Quais as consequências?
- Pode resultar que os emv dos coeficientes regressores sejam infinitos ou não existam.
- O problema de emv infinito também pode ocorrer em RL quando a variável preditora é contínua e ocorre a completa ou quasi-completa separação.
- Vejamos alguns exemplos a seguir.

Exemplo 2 (Agresti, pag.137)

```
> x <- c(10, 20, 30, 40, 60, 70, 80, 90); y <- c(0, 0, 0, 0, 1, 1, 1, 1)
> fit <- glm(y ~ x, family = binomial)
Warning message: glm.fit: fitted probabilities numerically 0 or 1 occurred
> summary(fit)

             Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  -118.158  296046.187        0         1
x               2.363    5805.939        0         1 # P-value for Wald test
---                                                  # of H0: beta=0
    Null deviance: 1.1090e+01  on 7  degrees of freedom
Residual deviance: 2.1827e-10  on 6  degrees of freedom # res. deviance = 0
                                                       # perfect fit
Number of Fisher Scoring iterations: 25 # very slow convergence
```

Exemplo 2 (Agresti, pag.137)

- Considerando apenas o valor-P, não rejeitaríamos a hipótese de $H_0 : \beta_1 = 0$. Neste caso, a conclusão seria de que x não influência y.
- Notamos também que os valores dos erros padrões são muito altos, indicando uma grande incerteza associada as estimativas.
- Na figura abaixo apresentamos os dados para entender o que esta ocorrendo.

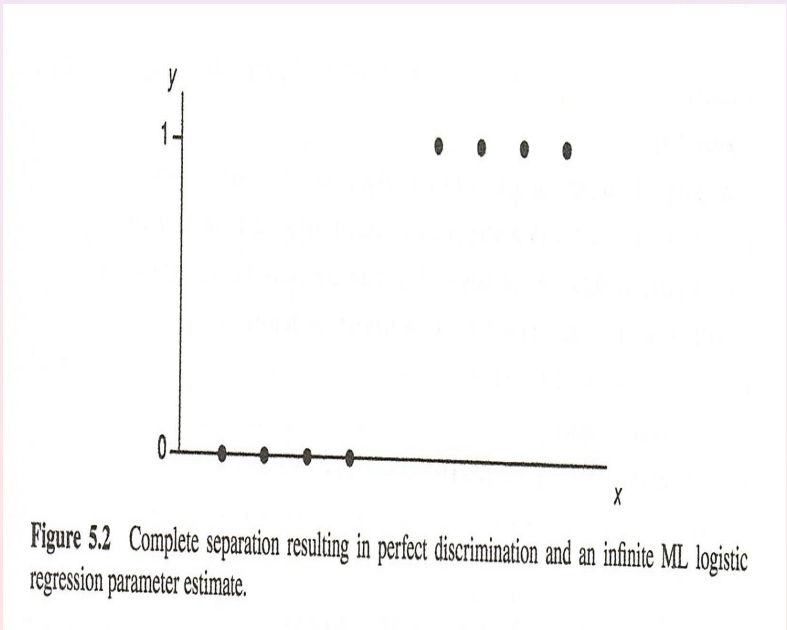

Figure 5.2 Complete separation resulting in perfect discrimination and an infinite ML logistic regression parameter estimate.

Exemplo 2 (Agresti, pag.137)

- Os dados apresentam uma separação completa, isto é, $y = 0$ se $x < x_c$ e $y = 1$ se $x > x_c$.
- Essa condição evidencia uma alta associação entre as duas variáveis (perfeita discriminação).
- Logo, a conclusão obtida via valor-P não é correta.
- Também não são confiáveis as estimativas pontuais obtidas.
- A real estimativa de máxima verossimilhança para β_1 é ∞ .
- A intuição para este resultado vem da observação do gráfico, que nos indica que o melhor ajuste é aquele dado com uma curva com declividade maior possível (paralela ao eixo vertical).

Exemplo 2 (Agresti, pag.137)

- Formalmente deve-se provar que a função de verossimilhança $L(\beta_0, \beta_1)$ cresce infinitamente.
- Note que o valor da *deviance* residual é muito baixa, indicando um bom ajuste.
- A diferença entre as funções desvios é aproximadamente 11.1 com 1 graus de liberdade (valor-P < 0.0009) . Rejeita-se $H_0 : \beta_1 = 0$. Forte evidência contra H_0.
- Há discordância entre os testes de Wald e RV. O teste da razão de verossimilhança é mais confiável.

EMV infinito: Tabelas 2×2

No próximo exemplo vamos tratar do problema quando a variável explicativa é categórica. Ilustrando o problema com tabelas 2×2 .

Covariável X	Resposta Y		Totais
	j=1	j=2	
i=1	n_{11}	n_{12}	n_{1+}
i=2	n_{21}	n_{22}	n_{2+}
Totais	n_{+1}	n_{+2}	n

EMV infinito: Tabelas 2×2

Lembre que em RL, $\exp\{\beta_1\}$ representa a razão de chances.
A estimativa de OR é dada por

$$\hat{OR} = \frac{n_{11} n_{22}}{n_{12} n_{21}}$$

Primeiro caso: $n_{11} = 0$ (ou $n_{22} = 0$) e n_{12}, n_{21} diferentes de zero.
Neste caso $\hat{OR} = 0$ e portanto $\hat{\beta}_1 = -\infty$.

Segundo caso: $n_{12} = 0$ (ou $n_{21} = 0$) e n_{11}, n_{22} diferentes de zero.
Neste caso $\hat{OR} = \infty$ e portanto $\hat{\beta}_1 = \infty$.

Esses dois casos definem tabelas de separação parcial.

EMV infinito: Tabelas 2×2

Terceiro caso: Tabelas de separação total .

$n_{11} = n_{22} = 0 \quad$ ou $\quad n_{12} = n_{21} = 0$

As estimativas de máxima verossimilhança são $-\infty$ e ∞, respectivamente.

Terceiro caso: $n_{11} = n_{21} = 0 \quad$ ou $\quad n_{12} = n_{22} = 0.$

Nestes casos, $\hat{OR} = \frac{0}{0}$ e o estimador de máxima verossimilhança não existe.

Exemplo 3 (Agresti, pag. 139)

Exemplo 3: Em um estudo sobre câncer endometrial com 79 pacientes foi investigado como y=grau de histologia (0=baixo , 1 = alto) se relaciona com três fatores de riscos: x_1= neovasculação (1=presente, 0=ausente) , x_2= índice de pulsatilidade da artéria uterina (variando de 0 a 49) e x_3= altura do endométrio (variando de 0.27 a 3.61).
Os dados apresentam a seguinte característica: todos os 13 pacientes com $x_1 = 1$ tiveram grau de histologia alto ($y = 1$).
Tabela relacionando y com x_1:

	y=0	y=1
$x_1 = 0$	49	17
$x_1 = 0$	0	13

Exemplo 2 (Agresti, pag. 139)

```
> fit <- glm(HG ~ NV + PI + EH, family=binomial, data=Endo)
> summary(fit)

            Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)   4.305       1.637    2.629   0.0086
NV           18.186    1715.751    0.011   0.9915 # true estimate = infinity
PI           -0.042       0.044   -0.952   0.3413
EH           -2.903       0.846   -3.433   0.0006
---

    Null deviance: 104.903  on 78  degrees of freedom
Residual deviance:  55.393  on 75  degrees of freedom
```

Exemplo 2 (Agresti, pag. 139)

- O modelo ajustado é $logito(p(x)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
- O valor real da estimativa de mv é $\hat{\beta}_1 = \infty$.
- Observamos na saída do **R** uma alto valor de erro padrão associado à x_1 (NV) . A estimativa pontual apresentada e o valor-P associados não devem ser consideradas.
- A estatística para testar $H_0 : \beta_1 = 0; \beta_2 = 0; \beta_3 = 0$ baseada na função desvio é $D_0 - D_1 = 104.90 - 55.39 = 49.51$ com 3 graus de liberdades ($valor - P < 0.000$). Rejeita-se H_0.

Exemplo 3 (Agresti, pag. 139)

- Testes individuais para cada parâmetros podem ser realizados usando a função *Anova*(*ajuste*) que compara as funções desvios. Por exemplo, para testar $H_0 : \beta_1 = 0$ obteve-se o valor-P = 0.00222, mostrando a significância da variável x_1 como desejado.

- No entanto, ainda não temos uma estimativa rasoável para β_1. O emv não é util para esta finalidade.

- O uso da inferência Bayesiana ou métodos de verossimilhança penalizadas podem solucionar este problema.

- Vamos apresentar a seguir a primeira solução (IB).

Análise Bayesiana em RL

- Primeiro deve-se estabelecer uma distribuição *a priori* para os coeficientes regressores $\beta = (\beta_0, \beta_1, \ldots, \beta_q)$.
- A distribuição *a posteriori* é obtida via fórmula de Bayes

$$f(\beta \mid y, x) \propto f(\beta) L(\beta).$$

- Toda inferência deve ser feita com base na distribuição *a posteriori* e expressa como probabilidades.
- Devido a forma da função de verossimilhança não obtemos uma expressão conhecida para $f(\beta \mid y, x)$ e deve-se considerar aproximações.
- Uma possibilidade é usar a aproximação normal para distribuição *a posteriori*. Outra possibilidade é aproximar via simulação estocástica.

Análise Bayesiana em RL

- No caso do uso de simulação, o mais comum é considerar métodos de Monte Carlo baseados em Cadeias de Markov (sigla em inglês MCMC).
- Existem vários software Bayesianos para fazer essa estimação. Em particular, o MCMCpack do **R** .
- Uma especificação comum para $f(\beta)$ é a normal multivariada centrada em zero com matriz diagonal (independentes) e parâmetro de precisão (1/variância) comum. Deste modo, é necessário apenas especificar a precisão que esta associado a sua certeza/incerteza a respeito dos parâmetros regressores serem significantes.
- Alternativamente, podemos especificar uma distribuição *a priori* para os logitos e a partir desta induzir uma distribuição para β (ver Das and Dey, 2007) .

Análise Bayesiana em RL : Exemplo 3 (Agresti, pag.141)

```
> library(MCMCpack) # b0 = prior mean, B0 = prior precision = 1/variance
> fitBayes <- MCMClogit(HG ~ NV2 + PI2 + EH2, mcmc=10000000, b0=0, B0=0.01,
+                       data=Endo) # prior var. = 1/0.01 = 100, std dev = 10
> summary(fitBayes)
1. Empirical mean and standard deviation: # posterior distribution
             Mean     SD
(Intercept)  3.215  2.560
NV2          9.120  5.097
PI2         -0.473  0.454
EH2         -2.138  0.593
2. Quantiles for each variable:
              2.5%     25%     50%     75%   97.5%
(Intercept) -0.342   1.271   2.722   4.687   9.346
NV2          2.109   5.234   8.128  12.048  21.343
PI2         -1.414  -0.767  -0.455  -0.159   0.366
EH2         -3.403  -2.515  -2.101  -1.722  -1.082
> mean(fitBayes[,2] < 0) # probability below 0 for 2nd model parameter (NV2)
[1] 0.000223
```

Análise Bayesiana em RL : Exemplo 3 (Agresti, pag.141)

- Para esta análise as variáveis explicativas foram padronizadas. A comparação desses resultados com os obtidos por mv só pode ser feita após um novo ajuste usando *glm* com as variáveis padronizadas.
- A variância *a priori* foi fixada em $1/B_0 = 100$.
- A saída apresenta as estimativas das médias, desvios padrões e quantis das distribuições *a posteriori* marginais de cada β_j.
- A saída também apresenta a estimativa da $P(\beta_1 \leq 0 \mid y, x) = 0.000223$.
- Portanto, com uma probabilidade muito alta $\beta_1 > 0$.

Análise Bayesiana em RL : Exemplo 3 (Agresti, pag.141)

- O intervalo de credibilidade 0.95 (caudas iguais) para β_1 é $(2.1, 21.3)$, bastante amplo indicando uma grande incerteza sobre o verdadeiro valor. Isso é devido as formas das funções de verossimilhança e *a priori (flat)*.

- Usando a média como estimativa pontual temos que a estimativa da razão de chances associada a x_1 =neovasculação é $\exp(9.12) = 9136$. Com o uso da mediana, este valor resulta em $\exp(8.128) = 3388$.

- A diferença entre média e mediana indica assimetria na distribuição *a posteriori* e o uso do intervalo simétrico pode não ser o mais adequado.

Análise Bayesiana em RL : Exemplo 3 (Agresti, pag.141)

Table 5.5 Results of Bayesian and frequentist fitting of models to the endometrial cancer data-set of Table 5.4 .

Analysis	$\hat{\beta}_1\,(SD)$	Interval[a]	$\hat{\beta}_2\,(SD)$	$\hat{\beta}_3\,(SD)$
ML	∞ (—)	(1.3, ∞)	−0.42 (0.44)	−1.92 (0.56)
Bayes, $\sigma = 10$	9.12 (5.10)	(2.1, 21.3)	−0.47 (0.45)	−2.14 (0.59)
Bayes, $\sigma = 1$	1.65 (0.69)	(0.3, 3.0)	−0.22 (0.33)	−1.77 (0.43)

[a]Profile-likelihood interval for ML and equal-tail posterior interval for Bayes.

Análise Bayesiana em RL : Exemplo 3 (Agresti, pag.141)

- A tabela apresenta a comparação das estimativas pontuais e intervalar de MV e Bayesianas com duas distribuições *a priori*.
- Notamos grandes diferenças na estimação do parâmetro β_1.
- Para os parâmetros β_2 e β_3, EMV e Bayes com priori vaga (baixa precisão) são bem similares.
- A última linha da tabela mostra a forte influência da escolha da distribuição *a priori* nas estimativas.
- A escolha *a priori* da $N(0, 1)$ indica a crença do pesquisador, antes da realização do experimento, de que os efeitos não devam ser tão extremos.
- Mais precisamente, por exemplo, com uma probabilidade muito alta (0.99) acredita-se que $|\beta_1| < 3$.