Estatística aplicada à epidemiologia II
Modelos para desfecho binário

Leo Bastos – leonardo.bastos@fiocruz.br

PROCC – Fundação Oswaldo Cruz

https://github.com/lsbastos/eae2

FIOCRUZ

Regressão logística múltipla

- O modelo logístico ~~multivariado~~ múltiplo
- Seja Y um desfecho binário, e $X_1, X_2, \ldots, X_p$, p variáveis explicativas.
- O modelo de regressão logística é dado por

$$Y_i \sim Bernoulli(\theta_i)$$

onde

$$logit(\theta_i) = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi}$$

- Exponencial dos coeficientes pode ser interpretado como razão de chances da variável associada ao coeficiente fixada todas as outras em um mesmo valores (independente qual for esse valor).

FIOCRUZ

A função de ligação *logit*

- Função *logit*

$$\log\left(\frac{\theta_i}{1-\theta_i}\right) = \boldsymbol{x}_i^T\boldsymbol{\beta} = \eta_i \Rightarrow \mu_i = \frac{e^{\eta_i}}{1+e^{\eta_i}}$$

- Suponha que dois cenários serão comparados, onde a única diferença entre eles é a variável x_k, que assume valor a em um cenário e b no outro.

$$\log\left(\frac{\theta_{(x_k=x)}}{1-\theta_{(x_k=x)}}\right) = \boldsymbol{x}_{-k}^T\boldsymbol{\beta}_k + x\beta_k \Rightarrow \frac{\theta_{(x_k=x)}}{1-\theta_{(x_k=x)}} = \exp\{\boldsymbol{x}_{-k}^T\boldsymbol{\beta}_k + x\beta_k\}$$

- Logo a razão de chances de interesse é

$$OR = \frac{\frac{\theta_{(x_k=a)}}{1-\theta_{(x_k=a)}}}{\frac{\theta_{(x_k=b)}}{1-\theta_{(x_k=b)}}} = \frac{\exp\{\boldsymbol{x}_{-k}^T\boldsymbol{\beta}_k + a\beta_k\}}{\exp\{\boldsymbol{x}_{-k}^T\boldsymbol{\beta}_k + b\beta_k\}} = \exp\{\beta_k(a-b)\}$$

- Se $a = 1$ e $b = 0$ então

$$OR = \exp\{\beta_1\}$$

FIOCRUZ

Exemplo: Sífilis em usuários de drogas

- Seja um estudo transversal que recrutou aleatoriamente 605 usuários de drogas (heavy drug users)
- Heavy drug user: É aquele usuário que: usou droga injetável nos últimos 6 meses e/ou usou droga da pasta base da cocaína pelo menos 25 vezes nos ultimos 6 meses.
- Tem-se o interesse em quantificar, entre outras coisas, a associação entre a infecção de sífilis com sexo e idade dos participantes do estudo.

FIOCRUZ

Exemplo: Sífilis em usuários de drogas

```
> # dados <- read.csv("Aula6_binary/DUsifilis.csv")
> dados <- read.csv("DUsifilis.csv")
> head(dados)
  sifilis      sexo idade  faixaetaria
1       0 masculino    46 35 a 49 anos
2       0 masculino    29 25 a 34 anos
3       0 masculino    44 35 a 49 anos
4       0 masculino    20 18 a 24 anos
5       0 masculino    26 25 a 34 anos
6       0 masculino    39 35 a 49 anos
```

FIOCRUZ

Primeiro modelo

- *Sifilis* $\sim$ *Sexo*

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.7975	0.1703	-4.68	0.0000
sexomasculino	-0.5443	0.2067	-2.63	0.0084

- $OR_{Sexo:Masc}$ =0.58. (Qual a interpretação?)

FIOCRUZ

Mudando a referência

- Eu gostaria de interpretar como aumento na OR, olhar a OR com a mulher no numerador.
- Solução rápida, inverter o sinal do β. Pois $OR_{Sexo:Masc} = 1/OR_{Sexo:Fem}$ (É mesmo?)
- E portanto, $OR_{Sexo:Fem} = 1/OR_{Sexo:Masc} = 1/e^{\beta} = e^{-\beta}$
- No exemplo: $OR_{Sexo:Fem} = 1/OR_{Sexo:Masc}$ =1.72
- Um outro caminho, mudar a categoria de base no R usando a função relevel().

FIOCRUZ

Controlando pela idade

- $Sifilis \sim Sexo + Idade$

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.0659	0.2463	-8.39	0.0000
sexofeminino	0.7422	0.2183	3.40	0.0007
faixaetaria25 a 34 anos	0.6171	0.2781	2.22	0.0265
faixaetaria35 a 49 anos	1.0353	0.2858	3.62	0.0003
faixaetaria50 anos ou mais	1.1607	0.3646	3.18	0.0015

- $OR_{Sexo:Fem}$ =2.1

FIOCRUZ

Analise da deviance

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL			604	659.34	
sexo	1	6.76	603	652.58	0.0093
faixaetaria	3	16.95	600	635.63	0.0007

FIOCRUZ

Efeitos e intervalos

	OR	2.5 %	97.5 %
sexo	1.72	1.15	2.58
sexo + faixaetaria	2.10	1.37	3.22

FIOCRUZ

Estatística de bondade de ajuste

- A deviance é uma medida de bondade de ajuste.
- Hosmer e Lemeshow (1980) e Lemeshow e Hosmer (1982) propuzeram uma medida de bondade de ajuste mais apropriada para dados binários.
- Ideia:
  - Particionar os possíveis valores de probabilidade em g grupos (Ex. decis).
  - Para cada grupo, contar quantos casos foram observados cujo valor predito deveria pertencer àquele grupo.
  - Dessa forma, tem-se um tabela $g \times 2$,
  - Se o valor predito representar as frequências das tabelas, então tem-se um bom modelo.

FIOCRUZ

Estatística de Bondade de ajuste Hosmer-Lemeshow

- Estatística de bondade de ajuste de Hosmer-Lemeshow, $\hat{C}$

$$\hat{C} = \sum_{k=1}^{g} \frac{(o_k - n'_k \bar{\pi}_k)^2}{n'_k \bar{\pi}_k (1 - \bar{\pi}_k)}$$

  onde g é o número de grupos, n'_k número de observações no grupo k, o_k número de casos observados no grupo k, e $\bar{\pi}_k$ é média das probabilidade estimada no grupo k.
- $\hat{C} \to \chi^2_{g-2}$

FIOCRUZ

Exemplo: Teste Hosmer-Lemeshow

Para o exemplo da sífilis, sexo e faixa etária geram 8 grupos, entao g = 8

	Grupo	O_k	n'_k	$\hat{\theta}_k$	E_k
1	masculino : 18 a 24 anos	13	105	0.11	11.81
2	masculino : 25 a 34 anos	24	143	0.19	27.20
3	masculino : 35 a 49 anos	37	138	0.26	36.29
4	masculino : 50 anos ou mais	18	58	0.29	16.70
5	feminino : 18 a 24 anos	11	58	0.21	12.19
6	feminino : 25 a 34 anos	25	66	0.33	21.80
7	feminino : 35 a 49 anos	13	32	0.43	13.71
8	feminino : 50 anos ou mais	1	5	0.46	2.30

$\hat{C} = 3.03 \rightarrow$ p-valor $= 0.806$

FIOCRUZ

Resíduos

- Quais são as suposições?
  - Independência
- A análise de resíduos para dados binários não é tão natural quanto para outros tipos de variáveis.
- Como o desfecho assume somente dois valores, a visualização de padrões é mais difícil.
- Análises para procurar de pontos de alavanca e *outliers* continuam válidas.

FIOCRUZ

Outliers ou pontos influentes

- Leverage (pontos de alavanca): h_{ii} (*hatvalues(modelo)*)

$$H = X^T(X^TX)^{-1}X^T$$

  Valores h_{ii} maiores que 2 ou 2 vezes p/n merecem uma olhada.
- Leave-one-out measures:
  - DFFIT: Diferença nos ajustes: $\hat{y}_i - \hat{y}_{i(-i)}$
  - DFBETA: Diferença no ajuste de cada coeficiente: $\hat{\beta}_k - \hat{\beta}_{k(-i)}$
  - Distância de Cook: Diferença no ajuste em todos coeficientes

FIOCRUZ

Outliers: Medidas de influência

```
> summary(influence.measures(m1Sex))
Potentially influential observations of
        glm(formula = sifilis ~ sexo + faixaetaria, family = binomial(),      data

    dfb.1_ dfb.sxfm dfb.fa3a dfb.fa4a dfb.faom dffit cov.r   cook.d hat
103 -0.05   0.12     0.01     0.02     0.16     0.21  1.03_*  0.01   0.03_*
104  0.05  -0.11    -0.01    -0.02    -0.14    -0.19  1.03_*  0.01   0.03_*
143  0.05  -0.11    -0.01    -0.02    -0.14    -0.19  1.03_*  0.01   0.03_*
384  0.05  -0.11    -0.01    -0.02    -0.14    -0.19  1.03_*  0.01   0.03_*
560  0.05  -0.11    -0.01    -0.02    -0.14    -0.19  1.03_*  0.01   0.03_*
```