Modelos Estatísticos I
Modelos para desfecho binário

Leo Bastos – leonardo.bastos@fiocruz.br

PROCC – Fundação Oswaldo Cruz

https://github.com/lsbastos/eae2

FIOCRUZ

Variáveis binárias

- É o desfecho mais comum em epidemiologia
- Óbito $\{S, N\}$; Acima de peso $\{IMC > 25, IMC \leq 25\}$; doente $\{S, N\}$; etc.
- A variável aleatória associada a um desfecho binário assume apenas dois valores numéricos, usualmente {0,1}.
- Exemplo:

$$Y = \begin{cases} 1, & \text{tem o desfecho de interesse,} \\ 0, & \text{não tem o desfecho de interesse.} \end{cases}$$

FIOCRUZ

Dados binários

Os dados binários podem ser apresentados em sua forma bruta:

	Y	X1	X2
1	0	1	Tratamento 1
2	1	1	Tratamento 1
3	1	0	Tratamento 1
4	1	1	Controle
5	1	0	Tratamento 2
6	1	1	Tratamento 2
7	0	0	Tratamento 2
8	0	0	Controle
9	1	0	Tratamento 1
10	1	0	Controle

FIOCRUZ

Dados binários

Ou na forma agregada:

	Y	n	X1	X2
1	3	10	0	Controle
2	6	30	0	Tratamento 1
3	12	45	0	Tratamento 2
4	3	20	1	Controle
5	20	50	1	Tratamento 1
6	1	15	1	Tratamento 2

FIOCRUZ

Modelo para dados binários

- Seja $Y_1, Y_2, \ldots, Y_n$ variáveis aleatórias independentes
- O desfecho é binário, i.e. $Y_i = \{0, 1\}$
- A distribuição usada para dados binários é a distirbuição Bernoulli

$$Y_i \sim Bern(\theta_i)$$

onde $\theta_i = P(Y_i = 1) \in (0, 1)$.

- Nosso interesse se encontra em tentar explicar a probabilidade do desfecho de interesse θ_i.
- Isso é feito de forma similar ao modelo linear

FIOCRUZ

Modelo para dados binários

- No modelo

$$Y_i \sim Bern(\theta_i)$$

onde $\theta_i = P(Y_i = 1) \in (0, 1)$.

- Tentamos explicar uma função da probabilidade θ_i a partir de uma variável exposição X da seguinte forma:

$$g(\theta_i) = \alpha + \beta x_i$$

onde $g(\cdot)$ é uma função que leva dos valores entre 0 e 1 para os reais, em matematiquês $g(\cdot) : (0, 1) \to \mathbb{R}$

- A função mais popular que cumpre esse papel é a função logit.

FIOCRUZ

Função logit

- A função logit é definida por

$$logit(\theta) = \log\left(\frac{\theta}{1-\theta}\right)$$

- O componente $\frac{\theta}{1-\theta}$ é conhecido por odds
- Veja a tabela abaixo:

	Prob	Odds	log(Odds)
1	0.10	0.11	-2.20
2	0.25	0.33	-1.10
3	0.50	1.00	0.00
4	0.75	3.00	1.10
5	0.90	9.00	2.20
6	0.99	99.00	4.60

FIOCRUZ

Voltando ao modelo

- Seja

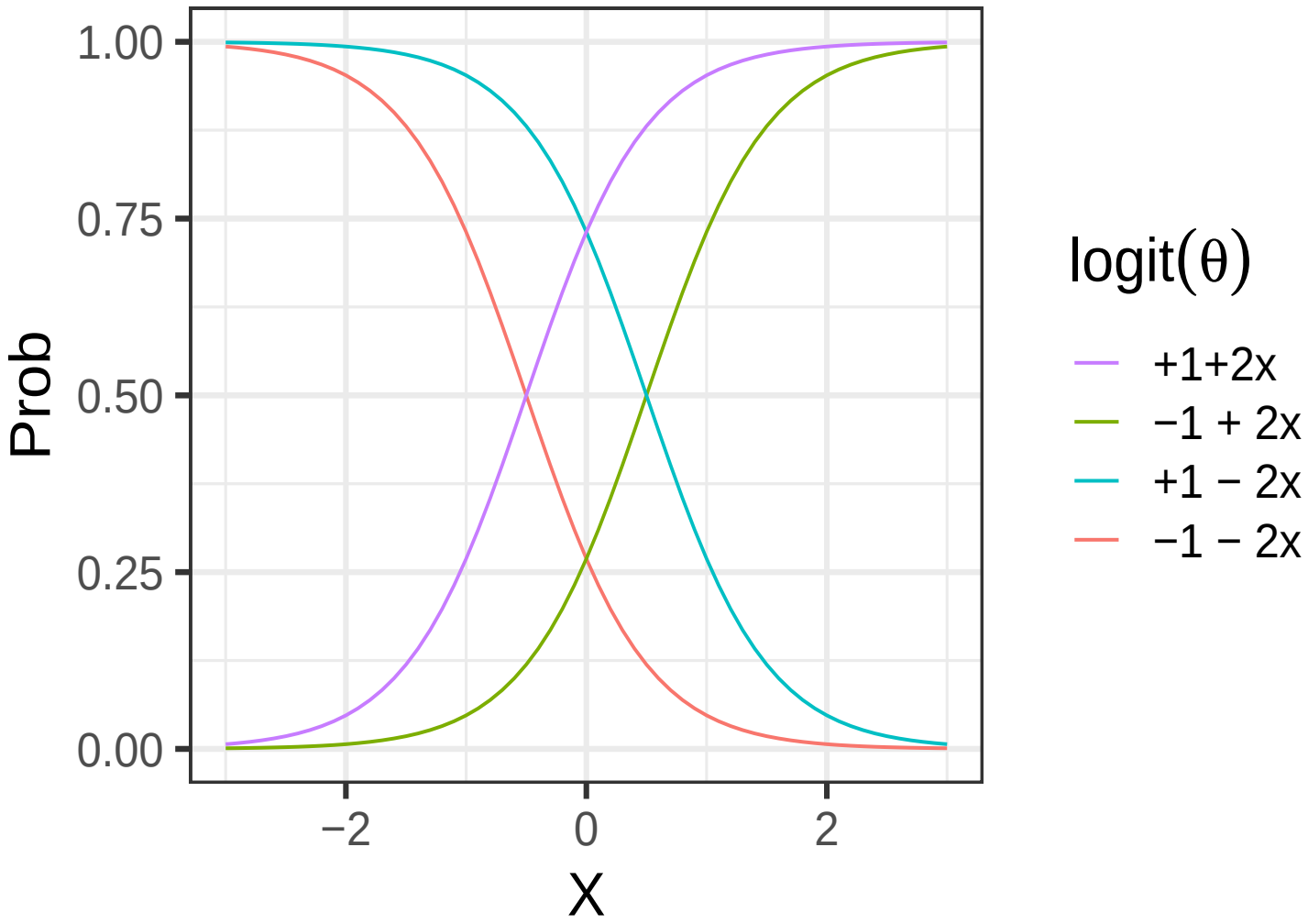
$$Y_i \sim Bern(\theta_i)$$

  onde $\theta_i = P(Y_i = 1) \in (0, 1)$.
- Suponha que temos uma exposição X
- Usando a função de ligação logit, temos que

$$logit(\theta_i) = \alpha + \beta x_i$$

FIOCRUZ

Função logit



FIOCRUZ

Função logit

- Na epidemiologia, uma medida de associação bastante popular é a razão de chances.
- Suponha uma exposição binária, $X = \{0, 1\}$, e $\theta_1 = P(Y = 1 \mid X = 1)$, e $\theta_0 = P(Y = 1 \mid X = 0)$
- A razão de chances, ou odds ratio, de X é dada por

$$OR_X = \frac{\frac{\theta_1}{1-\theta_1}}{\frac{\theta_0}{1-\theta_0}}$$

- Usando a função de ligação logit, temos que:

$$\log\left(\frac{\theta_x}{1-\theta_x}\right) = \alpha + \beta x \Rightarrow \frac{\theta_x}{1-\theta_x} = \exp\{\alpha + \beta x\}$$

FIOCRUZ

Função logistica

- Logo a razão de chances de interesse é

$$OR_X = \frac{\frac{\theta_1}{1-\theta_1}}{\frac{\theta_0}{1-\theta_0}} = \frac{\exp\{\alpha + \beta \times 1\}}{\exp\{\alpha + \beta \times 0\}} = \exp\{\beta\}$$

- Ou seja,

$$OR_X = \exp\{\beta_1\}$$

FIOCRUZ

Exemplo: Fumo na gestação versus baixo peso

- Em um estudo observacional sobre tabagismo na gestação realizado em um certo hospital, uma amostra de 189 puérperas foi aleatoriamente selecionada.
- Foi verificado se as mulheres fumaram ou não durante a gestação.
- o peso do recém nascido foi registrado, e categorizado como baixo peso sim ou não.
- Os dados observados foram

	Baixo peso ao nascer?	
Fumou durante a gestação?	Sim	Não
Sim	30	44
Não	29	86

FIOCRUZ

Os dados coletados

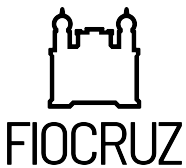
- A tabela de contingência Baixo peso ao nascer x mãe fumou durante gravidez

	Baixo peso ao nascer?	
Fumou durante a gestação?	Sim	Não
Sim	30	44
Não	29	86

- Calculando a OR:

$$OR = ?$$

- Essa OR é estatisticamente significativa?



Construindo o modelo

- Vamos construir o modelo desse exemplo
- Quem é o desfecho, Y?
  - Peso do recem nascido ($Y = 1$ se baixo peso)
- Quem é a exposição, X?
  - Fumo durante a gestação ($X = 1$ sim)
- Como é o modelo?
  - O modelo é dado por

$$Y \mid X = x \sim Bernoulli(\theta_x)$$

onde $logit(\theta_x) = \alpha + \beta x$

FIOCRUZ

No R

- No R, usamos a função *glm*

```
> saida <- glm(low ~ smoke,
+                family=binomial(link = 'logit'),
+                data=birth)
```

- Precisamos especificar a família, nesse caso é a Binomial (Lembrem-se que a $Bernoulli(\theta) = Binomial(1, \theta)$)
- A função de ligação default é a logit.

FIOCRUZ

Saída R: glm()

- A saída da função *glm* é bastante similar a saída da função *lm*

```
Call:  glm(formula = low ~ smoke, family = binomial(link = "logit"),
    data = birth)

Coefficients:
(Intercept)     smokesim
    -1.0871       0.7041

Degrees of Freedom: 188 Total (i.e. Null);  187 Residual
Null Deviance:            234.7
Residual Deviance: 229.8        AIC: 233.8
```

- Calculando a OR

```
> exp(0.7041)
[1] 2.022026
```

Exemplo

- Essa OR é significativa? (i.e. H_0: $OR = 1$ versus H_1: $OR \neq 1$)
- Como a $OR = exp(\beta)$, então o teste acima, é equivalente a testar H_0: $\beta = 0$ versus H_1: $\beta \neq 0$.
- Ou, alternativamente, podemos calcular um IC de 95% de confiança para a OR (ou para β)
- Um possível intervalo de 95% para a OR é dado pela exponencial do intervalo de 95% para β.
- Note que dessa forma o intervalo para a OR não é mais simétrico em torno da estimativa pontual, mas continua sendo um IC de 95% válido.

FIOCRUZ

Saidas no R

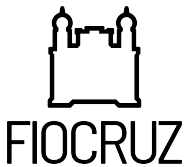
```
Call:
glm(formula = low ~ smoke, family = binomial(link = "logit"),
    data = birth)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.0871     0.2147  -5.062 4.14e-07 ***
smokesim      0.7041     0.3196   2.203   0.0276 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 229.80  on 187  degrees of freedom
AIC: 233.8

Number of Fisher Scoring iterations: 4
```



Saidas no R

- O IC para os coeficientes podemos ser calculados da seguinte forma

```
> cbind(saida$coef, confint(saida) )
                              2.5 %    97.5 %
(Intercept) -1.0870515 -1.5243118 -0.679205
smokesim     0.7040592  0.0786932  1.335154
```

- E o IC para o coeficiente é dado por

```
> exp(cbind(OR=saida$coef, confint(saida) ))[-1,]
      OR    2.5 %   97.5 %
2.021944 1.081872 3.800582
```

Função logistica

- Se for de interesse (e o delineamento do estudo permitir) podemos estimar a probabilidade do desfecho ocorrer sob condições específicas das variáveis explicativas.
- Ou seja,

$$P(Y_i = 1|X = x) = logit^{-1}(\alpha + \beta x)$$

- Para o modelo logistico:

$$P(Y_i = 1|X = x) = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}}$$

- Logo, previsões para essas probabilidades são dados por

$$\widehat{P}(Y_i = 1|X = x) = \frac{e^{\hat{\alpha}+\hat{\beta} x}}{1 + e^{\hat{\alpha}+\hat{\beta} x}}$$

FIOCRUZ

No exemplo

- No exemplo, podemos estimar a probabilidade de baixo para mulheres que fumam e para mulheres que não fumam durante a gravidez.
- Basta usar a função *predict*

```
> birth.pred <- data.frame(smoke = c("sim", "nao"))
> prev <- predict(saida, type = 'response',
+                 newdata = birth.pred)
> prev
        1         2
0.4054054 0.2521739
```

FIOCRUZ

Intervalo de confiança para a previsão

- Para evitar 'esquisitices' o intervalo de confiança para as previsões exigem um pouco mais de algebrismo.

```
> prev2 <- predict(saida, type = 'link',
+                  newdata = birth.pred,
+                  se.fit = T)
```

- Criando os intervalos

```
> birth.pred <- birth.pred %>%
+   bind_cols(
+     Prob = prev,
+     Link = prev2$fit,
+     Link.sd = prev2$se.fit
+   )
```

FIOCRUZ

Previsões

- A tabela estimada:

	smoke	Prob	Link	Link.sd
1	sim	0.405	-0.383	0.237
2	nao	0.252	-1.087	0.215

- IC para $\eta = \alpha + \beta x$

	smoke	Link	Link.sd	LI	LS
1	sim	-0.383	0.237	-0.847	0.081
2	nao	-1.087	0.215	-1.508	-0.666

- IC para $\theta_x = \mathit{logit}^{-1}(\alpha + \beta x)$

	smoke	Prob	LI	LS
1	sim	0.405	0.300	0.520
2	nao	0.252	0.181	0.339

FIOCRUZ