Modelos Lineares II

PROFESSOR: RAFAEL ERBISTI

Análise de dados binários e regressão logística

- Existem casos em que estamos interessados na relação entre uma resposta binária (ou dicotômica) e variáveis explicativas.

- A variável resposta pode ser, por exemplo, vivo ou morto, presente ou ausente, aprovado ou reprovado.

- Em geral, os termos para estas duas categorias são ”sucesso” e ”fracasso”.

Distribuições de probabilidade

- Podemos definir uma variável aleatória binária da forma

$$Z = \begin{cases} 1, & se\ o\ resultado\ \text{é}\ sucesso \\ 0, & se\ o\ resultado\ \text{é}\ fracasso \end{cases}$$

com $P(Z = 1) = \pi$ e $P(Z = 0) = 1 - \pi$. Então,

$$P(Z = z) = \pi^z(1-\pi)^{1-z}.$$

Distribuições de probabilidade

- Z tem distribuição de Bernoulli com parâmetro π, onde π é a probabilidade de sucesso.

- Temos que $E(Z) = \pi$ e $Var(Z) = (1 - \pi)$.

- Se temos n variáveis aleatórias $Z_1, \dots, Z_n$ independentes, tais que $P(Z_j = 1) = \pi_j$, então sua função de probabilidade conjunta é

$$\prod_{j=1}^{n} P(Z_j = z_j) = \exp\left[\sum_{j=1}^{n} z_j \log\left(\frac{\pi_j}{1-\pi_j}\right) + \sum_{j=1}^{n} \log(1-\pi_j)\right]$$

- É fácil ver que a distribuição de Bernoulli pertence à família exponencial.

Distribuições de probabilidade

- Para o caso em que os π_j's são todos iguais, podemos definir

$$Y = \sum_{j=1}^{n} Z_j$$

então Y é o número de sucessos em n ensaios e, portanto, $Y \sim Bin(n, \pi)$.

- Consideraremos o caso geral de uma amostra aleatória de tamanho N; $Y_1, \ldots, Y_N$ correspondendo ao número de sucessos em N diferentes subgrupos, de modo que $Y_i \sim Bin(n_i, \pi_i)$.

Distribuições de probabilidade

- Nesse contexto o log da função de verossimilhança é:

$$l(\pi_1, \dots, \pi_N; y_1, \dots, y_N) = \sum_{i=1}^{N} \left[y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + n_i \log(1-\pi_i) + \log\binom{n_i}{y_i} \right]$$

- Os valores observados podem ser dispostos na seguinte tabela:

	Subgrupos			
	1	2	…	N
Sucessos	Y_1	Y_2	…	Y_N
Fracassos	$n_1 - Y_1$	$n_2 - Y_2$	…	$n_N - Y_N$
Total	n_1	n_2	…	n_N

Tabela: Frequências para variáveis com distribuição binomial.

MLG

- O objetivo é descrever a proporção de sucessos, $P_i = Y_i/n_i$ em cada subgrupo em termos de covariáveis que caracterizam os subgrupos.

- Como $E(Y_i) = n_i\pi_i$ e então $E(P_i) = \pi_i$, modelamos as probabilidades π_i como

$$g(\pi_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}$$

onde $\boldsymbol{x}_i$ é o vetor de variáveis explicativas, $\boldsymbol{\beta}$ é o vetor de parâmetros e g é a função de ligação.

MLG

- O caso mais simples é o modelo linear

$$\pi = \boldsymbol{x}^T \boldsymbol{\beta}$$

- Esse modelo é utilizado algumas vezes na prática, apesar de apresentar a desvantagem dos valores ajustados $\hat{\pi} = \boldsymbol{x}^T \widehat{\boldsymbol{\beta}}$ poderem cair fora do intervalo $[0,1]$.

- Para garantir que π está restrito ao intervalo $[0,1]$, geralmente o modelamos através de uma função de distribuição acumulada

$$\pi = g^{-1}(\boldsymbol{x}^T \boldsymbol{\beta}) = \int_{-\infty}^{t} f(s)ds$$

onde $f(s) \geq 0$ e $\int_{-\infty}^{\infty} f(s)ds = 1$. $f(s)$ é chamada de **distribuição de tolerância**

Modelos de dose-resposta

- Um dos primeiros modelos de regressão para dados binomiais deve-se a Finney (1973).

- As respostas eram proporções de sucessos, por exemplo, proporção de animais mortos por diferentes níveis de doses de uma substância tóxica.

- O objetivo é descrever a probabilidade de "sucesso", π como função da dose, x, por exemplo, $g(\pi) = \beta_1 + \beta_2 x$.

Modelo probito

- Um dos modelos usados para dados binomiais é chamado de **modelo probito**.

- Neste modelo a distribuição normal é usada como distribuição de tolerância:

$$\pi = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left[-\frac{1}{2}\left(\frac{s-\mu}{\sigma}\right)^2\right] ds$$

$$= \Phi\left(\frac{x-\mu}{\sigma}\right)$$

onde Φ denota a função de distribuição acumulada da normal padrão.

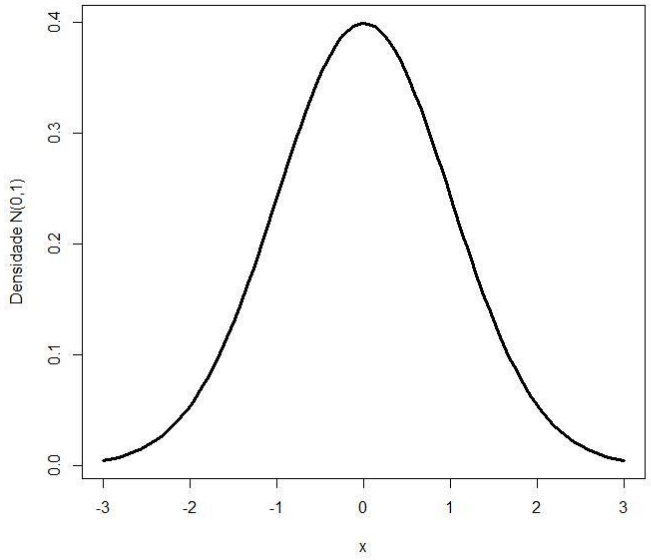
Modelo probito

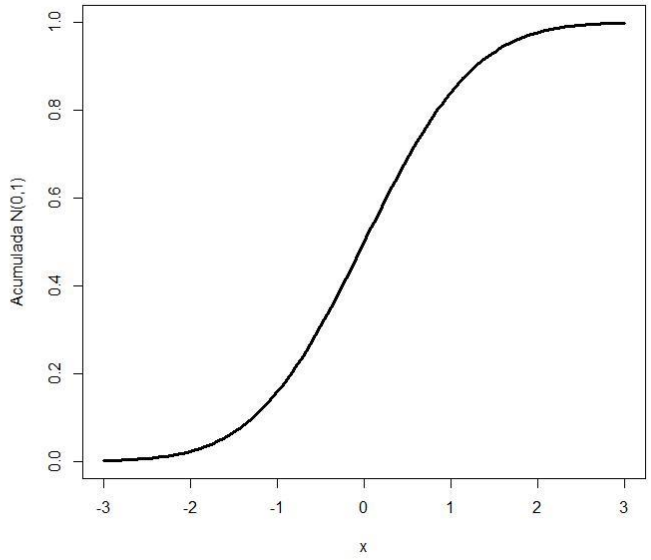
- Assim

$$\Phi^{-1}(\pi) = \beta_1 + \beta_2 x$$

onde $\beta_1 = -\mu/\sigma$ e $\beta_2 = 1/\sigma$, e a função de ligação é o inverso da função de distribuição acumulada da normal padrão Φ^{-1}.

Modelo probito





Densidade $f(s)$ $\pi(x)$

Figura: Distribuição $N(0,1)$.

Modelos logístico ou logito

- No **modelo logístico ou logito** a distribuição de tolerância é

$$f(s) = \frac{\beta_2 \exp(\beta_1 + \beta_2 s)}{[1 + \beta_2 \exp(\beta_1 + \beta_2 s)]^2}$$

de modo que

$$\pi = \int_{-\infty}^{x} f(s)ds = \frac{\exp(\beta_1 + \beta_2 x)}{1 + \exp(\beta_1 + \beta_2 x)},$$

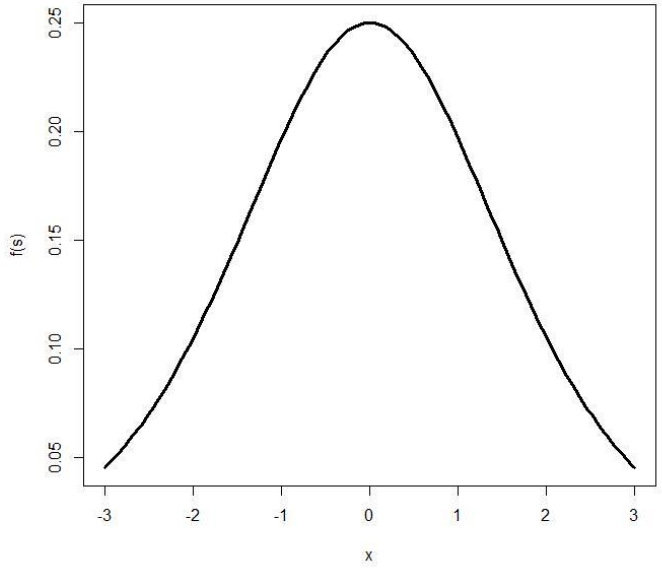
Modelo logístico ou logito

- Esse modelo resulta na seguinte função de ligação:

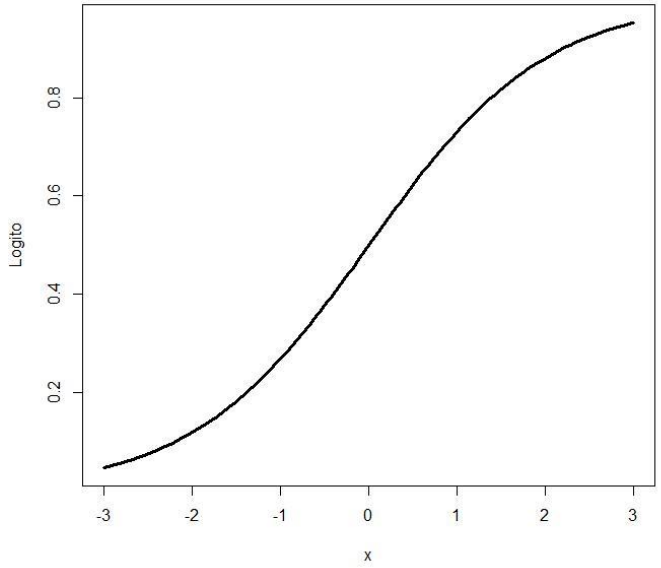
$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_1 + \beta_2 x$$

- Essa função de ligação é conhecida como **função logito**, e tem a interpretação natural como o log da razão de chances.

Modelo logístico ou logito



Densidade $f(s)$



$\pi(x)$

Figura: Distribuição logito

Modelo log-log complementar

- Um outro modelo é definido pela distribuição de valores extremos

$$f(s) = \beta_2 \exp[(\beta_1 + \beta_2 s) - \exp(\beta_1 + \beta_2 s)]$$

que leva a

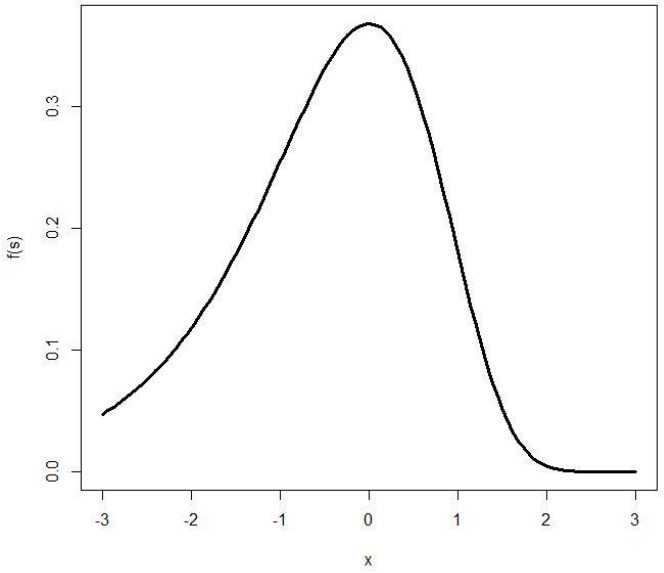
$$\pi = 1 - \exp[-\exp(\beta_1 + \beta_2 x)]$$

e

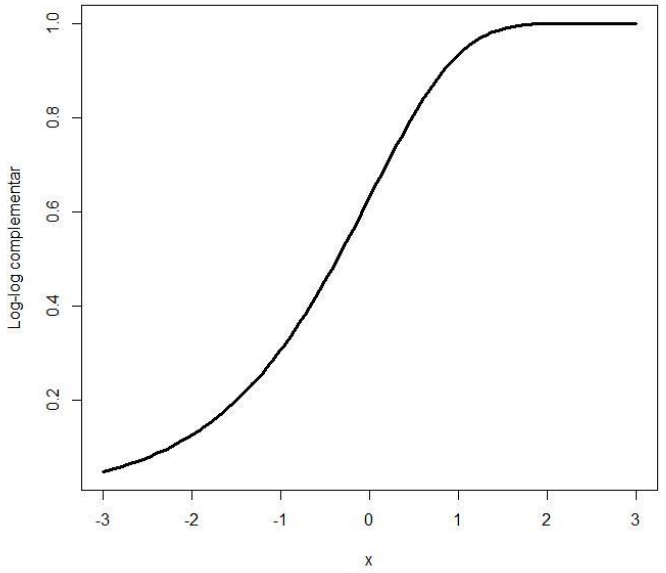
$$\log[-\log(1 - \pi)] = \beta_1 + \beta_2 x\,.$$

- Essa função de ligação é conhecida como **log-log complementar**.
- Ela é semelhante aos modelos probito e logito para valores de π próximos de 0,5, mas diferentes para valores de π próximos de 0 ou 1.

Modelo log-log complementar



Densidade $f(s)$



$\pi(x)$

Figura: Distribuição log-log complementar

Funções de ligação para dados com distribuição Bernoulli ou Binomial

Nome	Transformação	F.d.a	Distribuição
Logito	$logit(\pi) = \ln\left(\frac{\pi}{1-\pi}\right)$	$F(\eta) = \frac{\exp(\eta)}{1+\exp(\eta)}$	Logística
Probito	$probit(\pi) = \Phi^{-1}(\pi)$	$F(\eta) = \Phi(\eta)$	Normal
Log-log complementar	$cloglog(\pi) = \ln(-\ln(1-\pi))$	$F(\eta) = 1 - \exp(-\exp(\eta))$	Valor extremo

Observações

- As funções logística e probito são simétricas.

- A função log-log complementar não é simétrica.

- Quando a função de ligação é simétrica, o ajuste é o mesmo se consideramos $\pi = P(sucesso)$ ou $\pi = P(fracasso)$.

- As funções logito e probito são aproximadamente lineares para $0{,}1 < \pi < 0{,}9$, sendo difícil escolher qual das duas ligações deve ser utilizada baseando-se no bom ajuste.

Exemplo: ocorrência de sinistros

Considere o problema do Teste 1:

$$Y_i \sim Bernoulli(\pi_i), \qquad i = 1, \dots, N$$

$$\mathrm{g}(\pi_i) = \beta_0 + \beta_1 X_i$$

Iremos ajustar o modelo para o conjunto de dados e comprar os valores ajustados utilizando as funções de ligação logito e probito.

Exemplo: ocorrência de sinistros

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -6.005      3.114  -1.929   0.0538 .
x              2.193      1.007   2.178   0.0294 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 22.493  on 19  degrees of freedom
Residual deviance: 10.241  on 18  degrees of freedom
AIC: 14.241
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.5103     1.6793  -2.090   0.0366 *
x             1.2955     0.5377   2.409   0.0160 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 22.4934  on 19  degrees of freedom
Residual deviance:  9.9769  on 18  degrees of freedom
AIC: 13.977
```