GET00211 - Modelos Lineares 2

Rafael Erbisti

Instituto de Matemática e Estatística

Universidade Federal Fluminense

Aula 09

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".

- Podemos definir uma variável aleatória binária da forma:

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

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:
$$E(Z) = \pi \quad \text{e} \quad V(Z) = \pi(1-\pi)$$

- Se temos n variáveis aleatórias $Z_1, \ldots, Z_n$ independentes, tais que $P(Z_j = 1) = \pi_j$, então sua função de probabilidade conjunta é:
$$\prod_{j=1}^{n} P(Z_j = z_j) = \exp\left[\sum_{j=1}^{n} z_j \log \frac{\pi_j}{1-\pi_j} + \sum_{j=1}^{n} \log(1-\pi_j)\right]$$

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

Distribuições de probabilidade

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

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

então Y é o número de sucessos em n ensaios e, portanto:

$$Y \sim \text{Bin}(n, \pi)$$

- Consideraremos agora 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:

$$Y_i \sim \text{Bin}(n_i, \pi_i)$$

- Nesse contexto, o log da função de verossimilhança é:

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

- Os valores observados podem ser dispostos na seguinte tabela:

Table: Frequências para variáveis com distribuição binomial.

	Subgrupo 1	Subgrupo 2	...	Subgrupo 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

Modelos Lineares Generalizados

- O objetivo é descrever a proporção de sucessos em cada subgrupo:

$$P_i = \frac{Y_i}{n_i}$$

  em termos de covariáveis que caracterizam os subgrupos.

- Como:

$$E(Y_i) = n_i \pi_i \quad \text{e} \quad E(P_i) = \pi_i$$

  modelamos as probabilidades π_i como:

$$g(\pi_i) = \mathbf{x}_i' \boldsymbol{\beta}$$

- Onde:
  - $\mathbf{x}_i$ é o vetor de variáveis explicativas do subgrupo i,
  - $\boldsymbol{\beta}$ é o vetor de parâmetros a serem estimados,
  - g é a função de ligação.

- O caso mais simples é o modelo linear:

$$\pi = \mathbf{x}'\boldsymbol{\beta}$$

- Esse modelo é utilizado algumas vezes na prática, mas apresenta a desvantagem de que os valores ajustados:

$$\hat{\pi} = \mathbf{x}'\hat{\boldsymbol{\beta}}$$

  podem cair fora do intervalo $(0, 1)$.

- Para garantir que π esteja restrito ao intervalo $[0, 1]$, modelamos a probabilidade através de uma função de distribuição acumulada:

$$\pi = g^{-1}(\mathbf{x}'\boldsymbol{\beta}) = \int_{-\infty}^{t} f(s)\, ds$$

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

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:

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

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

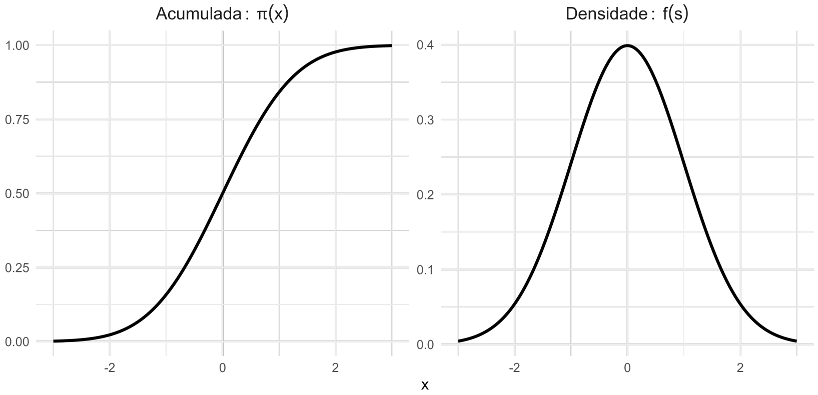
- Esse modelo resulta na seguinte função de ligação:

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

  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}.

- Essa função de ligação é conhecida como **função probito**.

Modelo Probito



Modelo Logístico (ou Logito)

- No modelo logístico (ou logito), a distribuição de tolerância é dada por:

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

- A função de distribuição acumulada é:

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

- Equivalentemente:

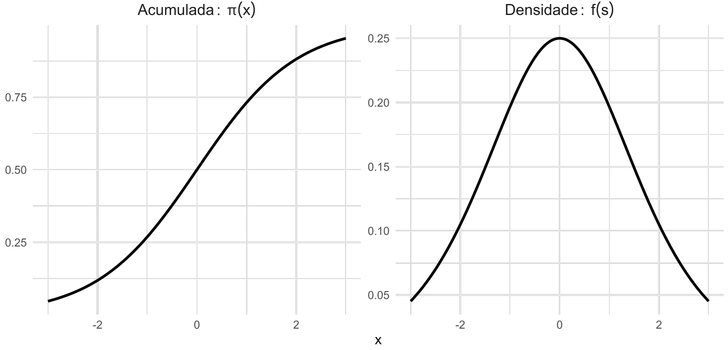
$$\pi(x) = \frac{1}{1 + \exp\big(-(\beta_1 + \beta_2 x)\big)}$$

Função de ligação do modelo logístico

- 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**.



Modelo log-log complementar

- Um outro modelo é definido pela distribuição de valores extremos:

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

- O que leva à função de distribuição acumulada:

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

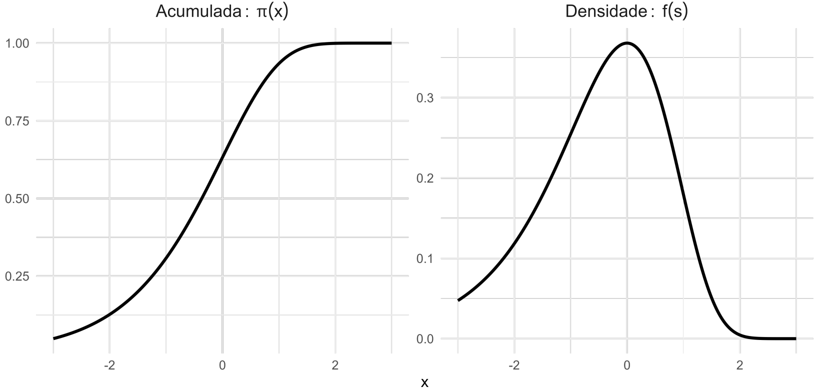
- A função de ligação correspondente é:

$$\log\big(-\log(1-\pi)\big) = \beta_1 + \beta_2 x$$

- 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 difere para valores de π próximos de 0 ou 1.

Modelo log-log complementar



Observações

- As funções logito 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 considerarmos $\pi = P(\text{sucesso})$ ou $\pi = P(\text{fracasso})$.
- As funções logito e probito são aproximadamente lineares para $0.1 < \pi < 0.9$, sendo difícil escolher entre elas apenas com base no bom ajuste.

Exemplo: ocorrência de sinistros

$$Y_i \sim \text{Bernoulli}(\pi_i), \quad i = 1, \ldots, n$$
$$g(\pi_i) = \beta_0 + \beta_1 X_i$$

- Ajustaremos o modelo para o conjunto de 20 segurados, com x sendo o valor do veículo e y a ocorrência ou não de sinistro.
- Compararemos os valores ajustados utilizando duas funções de ligação:

1. **Logito:**

$$g(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right)$$

2. **Probito:**

$$g(\pi_i) = \Phi^{-1}(\pi_i)$$

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
```