

ME720 - Modelos Lineares Generalizados

Parte 11 - Modelos para Dados Binários

Profa. **Larissa Avila Matos**

Modelos para Dados Binários

Para respostas binárias, assumimos geralmente uma distribuição binomial para o componente aleatório de um MLG.

Como vimos, a partir da sua representação na família exponencial com parâmetro de escala, o parâmetro natural da distribuição binomial é a *log-odds* (*log-chances*), ou seja, o *logito*,

$$\text{logito}(\pi) = \log \left(\frac{\pi}{1 - \pi} \right).$$

A função de ligação canônica para MLGs binomiais é o *logito*, para o qual o próprio modelo é chamado de **regressão logística**.

Este é o modelo mais importante para dados de resposta binária e é usado para uma ampla variedade de aplicações.

Regressão logística

Regressão logística

A regressão logística é caracterizada por questões de pesquisa com respostas binárias (sim/não ou sucesso/fracasso) ou respostas binomiais (número de sims ou sucessos em n ensaios).

Respostas binárias

As respostas binárias assumem apenas dois valores:

sucesso ($Y = 1$) ou fracasso ($Y = 0$).

Muitas vezes estamos interessados em modelar a probabilidade de sucesso π com base em um conjunto de covariáveis, embora algumas vezes desejemos usá-las para classificar uma observação futura como sucesso ou fracasso.

Exemplo: Os alunos com notas baixas têm maior probabilidade de beber em excesso? Assumindo que temos um conjunto de covariáveis para cada aluno.

Respostas Binomiais

As respostas binomiais são os números de sucessos em n ensaios idênticos e independentes com probabilidade constante π de sucesso.

Uma sequência de ensaios independentes com a mesma probabilidade de sucesso é chamada de processo de Bernoulli.

Assim como nas respostas binárias, nosso objetivo na modelagem de respostas binomiais é quantificar como a probabilidade de sucesso, π , está associada a covariáveis relevantes.

Distribuição logística

A função de distribuição logística foi proposta inicialmente para estudos demográficos, isto é, para estudos de crescimento populacional humano.

Seja Y variável aleatória contínua, dizemos que Y tem distribuição Logística com parâmetros de locação μ e de escala σ , se sua f.d.p. é dada por

$$f(y) = \frac{\exp\left(-\frac{y-\mu}{\sigma}\right)}{\sigma \left(1 + \exp\left(-\frac{y-\mu}{\sigma}\right)\right)^2}, \quad y, \mu \in \mathbb{R}, \sigma > 0,$$

e f.d.a. dada por

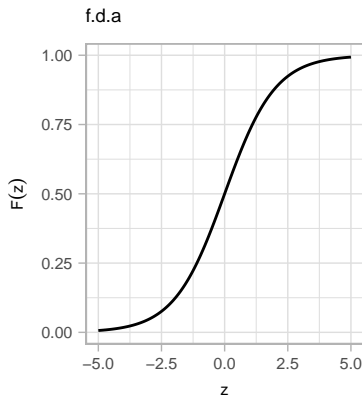
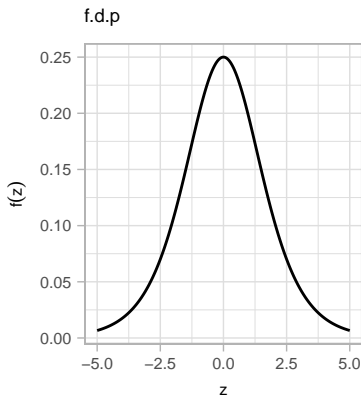
$$F(y) = \frac{1}{1 + \exp\left(-\frac{y-\mu}{\sigma}\right)}, \quad y, \mu \in \mathbb{R}, \sigma > 0.$$

Notação: $Y \sim \text{Logística}(\mu, \sigma)$.

Logística(0,1)

Seja $Z = \frac{X-\mu}{s}$, então a f.d.p. e a f.d.a de Z são dadas por

$$f(z) = \frac{e^{-z}}{(1 + e^{-z})^2} \quad \text{e} \quad F(z) = \frac{1}{1 + e^{-z}}; \quad z \in \mathbb{R}.$$



Pressupostos da regressão logística

- 1 **Resposta binária:** A variável resposta é dicotômica (duas respostas possíveis) ou a soma de respostas dicotômicas.
- 2 **Independência:** As observações devem ser independentes uma da outra.
- 3 **Estrutura de variação:** Por definição, a variação de uma variável aleatória binomial é $n\pi(1 - \pi)$, onde a variação é máxima quando $\pi = 0.5$
- 4 **Linearidade:** O log da razão de chances, $\log\left(\frac{\pi}{1-\pi}\right)$, deve ser uma função linear de \mathbf{X} .

Propriedades e Interpretações

Assumindo que Y_1, \dots, Y_N são proporções binomiais independentes, com $n_i Y_i \sim \text{Bin}(n_i, \pi_i)$.

O modelo para regressão logística possui duas formulações.

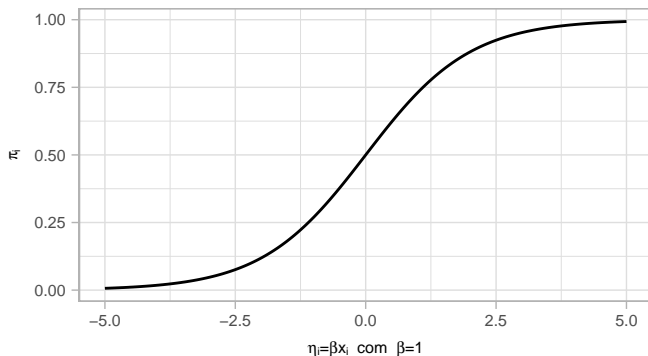
Modelo de regressão logística:

$$\pi_i = \frac{\exp(\sum_{j=1}^p \beta_j x_{ij})}{1 + \exp(\sum_{j=1}^p \beta_j x_{ij})} \quad \text{ou}$$

$$\text{logito}(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \sum_{j=1}^p \beta_j x_{ij},$$

onde π_i é a proporção da proporção de sucessos de Y_i .

Para uma única covariável x quantitativa com $\beta > 0$, a curva para π_i tem a forma da f.d.a de uma distribuição logística.



Como a densidade logística é simétrica, à medida que x_i muda, π_i se aproxima de 1 na mesma taxa que se aproxima de 0.

Com várias variáveis explicativas, uma vez que

$$1 - \pi_i = \left[1 + \exp\left(\sum_j \beta_j x_{ij}\right) \right]^{-1},$$

π_i é monótona em cada variável explicativa de acordo com o sinal de seus coeficientes. A taxa de subida ou descida aumenta à medida que $|\beta_j|$ aumenta.

Quando $\beta_j = 0$, Y é condicionalmente independente de x_j , dadas as outras variáveis explicativas.

Como interpretamos a magnitude de β_j ?

Temos que,

$$\frac{\partial \pi_i}{\partial x_{ij}} = \beta_j \frac{\exp(\sum_{j=1}^p \beta_j x_{ij})}{\left[1 + \exp(\sum_{j=1}^p \beta_j x_{ij})\right]^2} = \beta_j \pi_i (1 - \pi_i)$$

A inclinação é mais íngreme (e igual a $\frac{\beta_j}{4}$) no valor de x_{ij} para o qual $\pi_i = 1/2$, e a inclinação diminui para 0 quando π_i se move para 0 ou 1.

Como interpretamos β_j para uma variável explicativa qualitativa?

A interpretação dos parâmetros de um modelo de regressão logística é obtida comparando a probabilidade de sucesso com a probabilidade de fracasso, usando a função de razão de chances (RC). Essa função é obtida a partir da função de chances (odds),

$$\begin{aligned} g(\mathbf{x}) &= \frac{\pi_i(\mathbf{x})}{[1 - \pi_i(\mathbf{x})]} = \frac{\frac{e^{\sum_j \beta_j x_{ij}}}{1 + e^{\sum_j \beta_j x_{ij}}}}{1 - \frac{e^{\sum_j \beta_j x_{ij}}}{1 + e^{\sum_j \beta_j x_{ij}}}} = \frac{\frac{e^{\sum_j \beta_j x_{ij}}}{1 + e^{\sum_j \beta_j x_{ij}}}}{\frac{1}{1 + e^{\sum_j \beta_j x_{ij}}}} \\ &= e^{\sum_j \beta_j x_{ij}} = e^{\eta_i}. \end{aligned}$$

Exemplo:

- Se $\pi = 0,50$, a chance de ocorrência do evento será de 1 (1 para 1).
- Se $\pi = 0,75$, a chance de ocorrência do evento será de 3 (3 para 1).

Considerando uma única covariável x quantitativa, temos

$$g(x) = \frac{\pi_i(x)}{[1 - \pi_i(x)]} = e^{\beta_0 + \beta_1 x_i}.$$

Assim, ao tomarmos dois valores distintos da variável explicativa, x_j e x_{j+1} , obtemos

$$RC = \frac{g(x_{j+1})}{g(x_j)} = \frac{e^{\beta_0 + \beta_1 x_{j+1}}}{e^{\beta_0 + \beta_1 x_j}} = e^{\beta_1 (x_{j+1} - x_j)}.$$

Temos ainda que,

$$\begin{aligned}\log(RC) &= \log \left[\frac{g(x_{j+1})}{g(x_j)} \right] = \log [g(x_{j+1})] - \log [g(x_j)] \\ &= \beta_0 + \beta_1 x_{j+1} - \beta_0 - \beta_1 x_j = \beta_1 (x_{j+1} - x_j).\end{aligned}$$

Fazendo $x_{j+1} - x_j = 1$ unidade, então

$$\log(RC) = \log(e^{\beta_1}) = \beta_1.$$

Assim, temos o quão provável o resultado ocorrerá entre os indivíduos x_{j+1} em relação aos indivíduos x_j , portanto

$$\text{se } \beta_1 > 0 \quad \Rightarrow \quad RC > 1 \quad \Rightarrow \quad \pi(x_{j+1}) > \pi(x_j)$$

$$\text{se } \beta_1 < 0 \quad \Rightarrow \quad RC < 1 \quad \Rightarrow \quad \pi(x_{j+1}) < \pi(x_j).$$

Exemplo Livro Faraway: Desastre do Challenger

Em janeiro de 1986, o ônibus espacial Challenger explodiu logo após o lançamento.

Uma investigação foi iniciada para descobrir a causa do acidente; e a atenção concentrou-se nas vedações dos anéis de borracha (*O-rings*) nos propulsores do foguete.

Em temperaturas mais baixas, a borracha se torna mais quebradiça e é um selante menos eficaz.

No momento do lançamento, a temperatura era de 31°F.

Poderia ter sido prevista a falha dos *O-rings*?

Nas 23 missões anteriores de ônibus espaciais para as quais existem dados, foram registradas algumas evidências de danos causados por sopro e erosão em alguns *O-rings*.

Cada lançadeira tinha dois boosters, cada um com três *O-rings*, ou seja, seis *O-rings* por missão.

Para cada missão, sabemos o número de *O-rings* que mostram algum tipo de dano e a temperatura do lançamento.

Dados

```
library(faraway)
data(orings)
head(orings)
```

	temp	damage
1	53	5
2	57	1
3	58	1
4	63	1
5	66	0
6	67	0

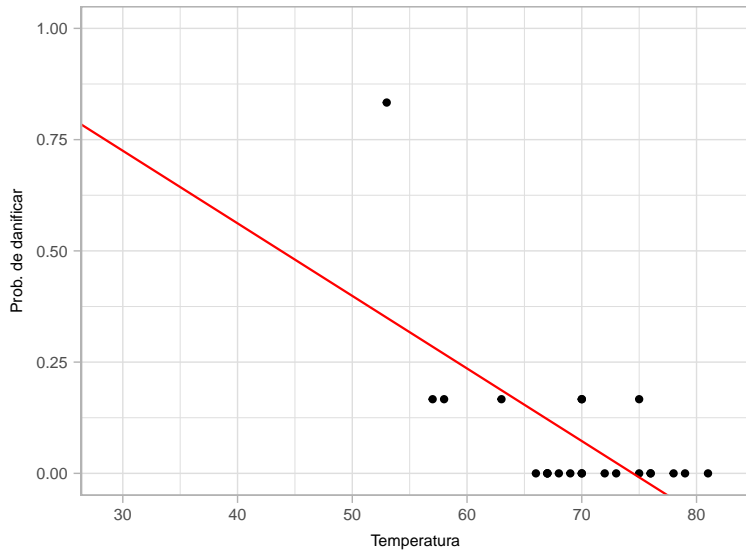
Resumindo os dados

```
dim(orings)
```

```
[1] 23  2
```

```
summary(orings)
```

temp		damage	
Min.	:53.00	Min.	:0.0000
1st Qu.:	:67.00	1st Qu.:	:0.0000
Median	:70.00	Median	:0.0000
Mean	:69.57	Mean	:0.4783
3rd Qu.:	:75.00	3rd Qu.:	:1.0000
Max.	:81.00	Max.	:5.0000



```
fit.logit <- glm(cbind(damage,6-damage) ~ temp, family=binomial, orings)
# cbind(damage,6-damage) 1a.coluna # de sucessos e 2.coluna # de fracassos
summary(fit.logit)
```

Call:

```
glm(formula = cbind(damage, 6 - damage) ~ temp, family = binomial,
     data = orings)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-0.9529	-0.7345	-0.4393	-0.2079	1.9565

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	11.66299	3.29626	3.538	0.000403 ***
temp	-0.21623	0.05318	-4.066	4.78e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 38.898 on 22 degrees of freedom
 Residual deviance: 16.912 on 21 degrees of freedom
 AIC: 33.675

Number of Fisher Scoring iterations: 6

Da estimativa dos parâmetros temos que

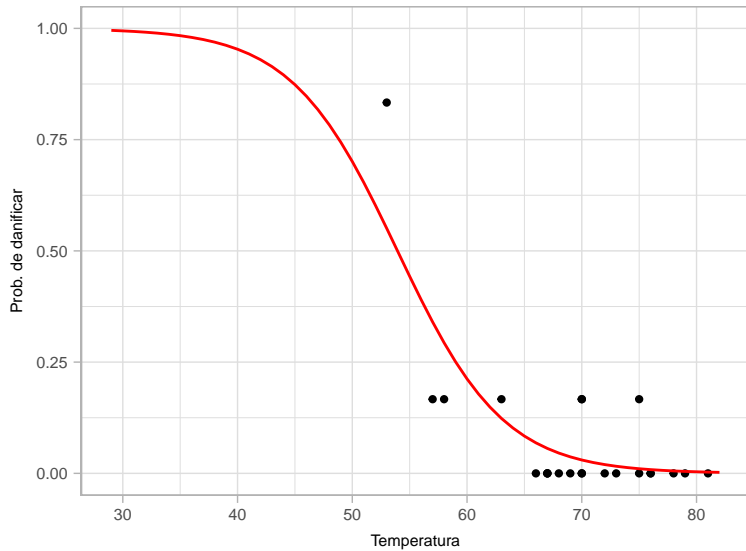
$$RC(\hat{\beta}_1) = e^{-0,21623} = 0,80555.$$

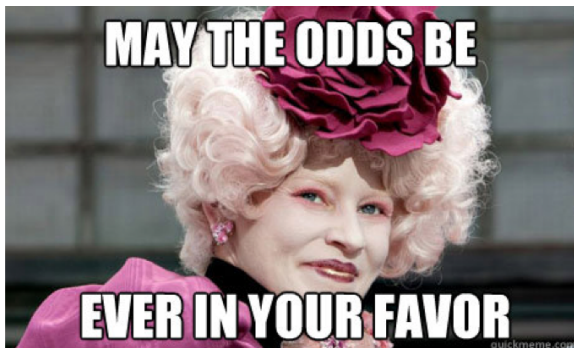
Como a RC é menor que 1, a probabilidade de danificar o *O-rings* tende a diminuir quando aumenta a temperatura do lançamento.

A chance de danificar o *O-rings* é 1,241388 ($e^{0,21623}$) vezes maior quando diminuimos uma unidade na temperatura.

```
cbind(oring$temp,fit.logit$fitted.values)
```

	[,1]	[,2]
1	53	0.550478817
2	57	0.340216592
3	58	0.293475686
4	63	0.123496147
5	66	0.068597710
6	67	0.056005745
7	67	0.056005745
8	67	0.056005745
9	68	0.045612000
10	69	0.037071413
11	70	0.030079600
12	70	0.030079600
13	70	0.030079600
14	70	0.030079600
15	72	0.019727169
16	73	0.015952356
17	75	0.010409884
18	75	0.010409884
19	76	0.008402660
20	76	0.008402660
21	78	0.005468670
22	79	0.004409961
23	81	0.002866088





Variável Binária

Vamos considerar variáveis explicativas assumindo valores discretos 0 ou 1.

Por exemplo, considere uma única variável explicativa x com níveis A ou B.

Para estimar os parâmetros do modelo, estas variáveis são substituídas por valores numéricos, por exemplo 0 para o nível A e 1 para o nível B. Então,

	$x=1$	$x=0$
$y=1$	$\pi(1) = \frac{e^{\beta_0 + \beta_1}}{1 + e^{\beta_0 + \beta_1}}$	$\pi(0) = \frac{e^{\beta_0}}{1 + e^{\beta_0}}$
$y=0$	$1 - \pi(1) = \frac{1}{1 + e^{\beta_0 + \beta_1}}$	$1 - \pi(0) = \frac{1}{1 + e^{\beta_0}}$
Total	1	1

Através dos valores da Tabela acima, a razão de chances será interpretada por:

$$RC = \frac{\frac{\pi(1)}{1 - \pi(1)}}{\frac{\pi(0)}{1 - \pi(0)}} = \frac{\frac{e^{\beta_0 + \beta_1}}{1 + e^{\beta_0 + \beta_1}}}{\frac{e^{\beta_0}}{1 + e^{\beta_0}}} = \exp(\beta_1)$$

Portanto, temos que

$$RC > 1 \quad \Rightarrow \quad \frac{\pi_1}{1 - \pi_1} > \frac{\pi_0}{1 - \pi_0} \quad \Rightarrow \quad \pi_B > \pi_A$$

$$RC < 1 \quad \Rightarrow \quad \frac{\pi_1}{1 - \pi_1} < \frac{\pi_0}{1 - \pi_0} \quad \Rightarrow \quad \pi_B < \pi_A.$$

Vimos que $\widehat{RC} = e^{\hat{\beta}_1}$. Mas esse resultado é válido apenas quando a variável independente é codificada em 0 ou 1.

A estimativa do log da RC para uma variável independente com dois níveis a e b é

$$\begin{aligned}\log[\widehat{RC}(a, b)] &= \hat{g}(x = a) - \hat{g}(x = b) \\ &= (\hat{\beta}_0 + \hat{\beta}_1 \cdot a) - (\hat{\beta}_0 + \hat{\beta}_1 \cdot b) = \hat{\beta}_1(a - b).\end{aligned}$$

Assim, a estimativa da RC é dada por:

$$\widehat{RC}(a, b) = e^{\hat{\beta}_1(a-b)}.$$

Inferência para os parâmetros do modelo

Equações de verossimilhança para um MLG

O sistema de equações de verossimilhança para um MLG é dada por

$$\begin{aligned}\frac{\partial \ell(\boldsymbol{\beta}, \phi)}{\partial \beta_j} &= \sum_{i=1}^n \frac{(y_i - \mu_i) x_{ij}}{\text{Var}(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0, \quad j = 1, 2, \dots, p, \quad \text{e} \\ \frac{\partial \ell(\boldsymbol{\beta}, \phi)}{\partial \phi} &= 0,\end{aligned}$$

onde $\eta_i = \sum_{j=1}^p \beta_j x_{ij} = g(\mu_i)$ para uma função de ligação g .

Para um MLG com dados binários, $n_i \sim \text{Bin}(n_i, \pi_i)$, $i = 1, \dots, N$, temos

$$\begin{aligned}\pi_i &= \mu_i \\ &= F\left(\sum_{j=1}^p \beta_j x_{ij}\right) = F(\eta_i),\end{aligned}$$

para alguma f.d.a padrão F .

Então,

$$\frac{\partial \mu_i}{\partial \eta_i} = f(\eta_i), \quad \text{onde } f \text{ é a f.d.p de } F.$$

Sabemos que

$$\mathbb{E}(Y_i) = \mathbb{E}\left(\frac{n_i Y_i}{n_i}\right) = \pi_i \quad \text{e} \quad \text{Var}(Y_i) = \text{Var}\left(\frac{n_i Y_i}{n_i}\right) = \frac{\pi_i(1 - \pi_i)}{n_i}.$$

Portanto, o sistema de equações de verossimilhança é dada por

$$\sum_{i=1}^N \frac{n_i (y_i - \pi_i) x_{ij}}{\pi_i(1 - \pi_i)} f(\eta_i) = 0, \quad j = 1, 2, \dots, p,$$

ou seja, em termos de β ,

$$\sum_{i=1}^N \frac{n_i \left(y_i - F(\sum_j \beta_j x_{ij}) \right) x_{ij}}{F(\sum_j \beta_j x_{ij})_i (1 - F(\sum_j \beta_j x_{ij}))} f \left(\sum_j \beta_j x_{ij} \right) = 0, \quad j = 1, 2, \dots, p.$$

Regressão logística

Para a regressão logística para dados binários, como vimos anteriormente,

$$f(z) = \frac{e^{-z}}{(1 + e^{-z})^2} \quad \text{e} \quad F(z) = \frac{1}{1 + e^{-z}} = F(z)(1 - F(z)); \quad z \in \mathbb{R}.$$

O sistema de equações de verossimilhança simplifica em

$$\sum_{i=1}^N \frac{n_i (y_i - \pi_i) x_{ij}}{F(\eta_i)(1 - F(\eta_i))} F(\eta_i)(1 - F(\eta_i)) = 0$$
$$\Rightarrow \sum_{i=1}^N n_i (y_i - \pi_i) x_{ij} = 0, \quad j = 1, 2, \dots, p.$$

Ou equivalentemente,

$$\sum_{i=1}^N n_i \left(y_i - F \left(\sum_j \beta_j x_{ij} \right) \right) x_{ij} = 0, \quad j = 1, 2, \dots, p.$$

Matriz de informação esperada

Como vimos, a matriz de informação esperada para um MLG é dada por

$$\mathbf{I} = \mathbf{X}'\mathbf{W}\mathbf{X},$$

com \mathbf{W} sendo uma matriz diagonal com elementos dados por

$$w_i = \frac{(\partial\mu_i/\partial\eta_i)^2}{\text{Var}(Y_i)}.$$

No caso da regressão logística, $\eta_i = \log\left(\frac{\pi_i}{1-\pi_i}\right)$, então temos que

$$\frac{\partial\eta_i}{\partial\pi_i} = \frac{1}{\pi_i(1-\pi_i)} \quad \Rightarrow \quad \frac{\partial\pi_i}{\partial\eta_i} = \frac{\partial\mu_i}{\partial\eta_i} = \pi_i(1-\pi_i)$$

$$\text{e } \text{Var}(Y_i) = \frac{\pi_i(1-\pi_i)}{n_i}.$$

Então,

$$\mathbf{I} = \mathbf{X}'\text{Diag}(n_i\pi_i(1-\pi_i))\mathbf{X},$$

Portanto, a matriz de covariância de $\hat{\beta}$ é dada pela inversa da matriz de informação \mathbf{I} .

Ajuste do modelo (regressão logística)

O Método de Newton-Raphson é equivalente ao Escore de Fisher, uma vez que o logito é a função de ligação canônica.

Sabemos que $\mathbf{H} = \mathbf{X}'\text{Diag}(n_i\pi_i(1 - \pi_i))\mathbf{X}$ e $\mathbf{u}^{(t)} = \mathbf{X}'(\mathbf{s} - \boldsymbol{\mu})$.

Se considerarmos $s_i = n_i y_i$, no passo t , o processo iterativo ($t = 0, 1, 2, \dots$) tem a seguinte aproximação

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - (\mathbf{H}^{(t)})^{-1}\mathbf{u}^{(t)},$$

onde $\mu_i^{(t)} = n_i\pi_i^{(t)}$ e

$$\pi_i^{(t)} = \frac{\exp\left(\sum_j \beta_j^{(t)} x_{ij}\right)}{1 + \exp\left(\sum_j \beta_j^{(t)} x_{ij}\right)}.$$

O procedimento descrito é repetido até que mudanças em $\ell(\boldsymbol{\beta}^{(t)})$ entre ciclos sucessivos são suficientemente pequenas.

Desvio e Estatística de Pearson para Bondade de Ajuste

Para a regressão logística o desvio é a estatística do teste da razão de verossimilhanças, comparando o modelo com o modelo saturado, dada por

$$\begin{aligned} LR &= -2 \log \left[\frac{\prod_{i=1}^N \hat{\pi}_i^{n_i y_i} (1 - \hat{\pi}_i)^{n_i - n_i y_i}}{\prod_{i=1}^N y_i^{n_i y_i} (1 - y_i)^{n_i - n_i y_i}} \right] \\ &= -2 \log \left[\prod_{i=1}^N \left(\frac{\hat{\pi}_i}{y_i} \right)^{n_i y_i} \left(\frac{1 - \hat{\pi}_i}{1 - y_i} \right)^{n_i - n_i y_i} \right] \\ &= 2 \sum_{i=1}^N n_i y_i \log \left(\frac{y_i}{\hat{\pi}_i} \right) + 2 \sum_{i=1}^N (n_i - n_i y_i) \log \left(\frac{1 - y_i}{1 - \hat{\pi}_i} \right) \\ &= 2 \sum_{i=1}^N n_i y_i \log \left(\frac{n_i y_i}{n_i \hat{\pi}_i} \right) + 2 \sum_{i=1}^N (n_i - n_i y_i) \log \left(\frac{n_i - n_i y_i}{n_i - n_i \hat{\pi}_i} \right), \end{aligned}$$

onde $n_i y_i$ é o número de sucessos e $(n_i - n_i y_i)$ é o número de fracassos, $i = 1, \dots, N$.

Portanto, o desvio é uma soma sob os $2N$ sucessos e fracassos totais nas N configurações, tendo a forma

$$D(\mathbf{y}; \hat{\boldsymbol{\mu}}) = 2 \sum \text{observado} \times \log \left(\frac{\text{observado}}{\text{ajustado}} \right).$$

Exercício: Mostre que o deavio difere quando temos dados agrupados e dados não agrupados, onde

- dados agrupados: $\mathbf{n} = (n_1, n_2, \dots, n_N)$,
- dados não agrupados: $\mathbf{n} = (1, 1, \dots, 1)$.

Para dados agrupados, a estatística de Pearson é dada por

$$\begin{aligned} X^2 &= 2 \sum \frac{(\text{observado} - \text{ajustado})^2}{\text{ajustado}} = \sum_{i=1}^N = \frac{(n_i y_i - n_i \hat{\pi}_i)^2}{n_i \hat{\pi}_i} \\ &= \frac{(y_i - \hat{\pi}_i)^2}{\hat{\pi}_i(1 - \hat{\pi}_i)/n_i}. \end{aligned}$$

Para mais detalhes ler seção 5.5.2 do livro texto.

Para a regressão logística, temos que os resíduos de Pearson é dada por

$$e_i = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}} = \frac{y_i - \hat{\pi}_i}{\widehat{\text{Var}}(Y_i)} = \frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1 - \hat{\pi}_i)/n_i}};$$

e o resíduo desvio padronizado é dado por

$$t_{d_i} = \frac{e_i}{\sqrt{(1 - \hat{h}_{ii})}} = \frac{y_i - \hat{\pi}_i}{\sqrt{\left[\hat{\pi}_i(1 - \hat{\pi}_i)(1 - \hat{h}_{ii})\right]/n_i}},$$

onde $\widehat{H}_W = \widehat{W}^{1/2} \mathbf{X}(\mathbf{X}'\widehat{W}\mathbf{X})^{-1} \mathbf{X}'\widehat{W}^{1/2}$, com $\widehat{w}_{ii} = n_i \hat{\pi}_i(1 - \hat{\pi}_i)$.

Obs.: $X^2 = \sum_{i=1}^N e_i^2$.

Bandas de confiança

Uma banda assintótica de confiança de coeficiente $1 - \alpha$ pode ser construída para $\pi(\mathbf{z})$, $\forall \mathbf{z} \in \mathbb{R}^p$ (ver também Piegorsch e Casella, 1988).

Assintoticamente $\hat{\beta} - \beta \sim N_p(\mathbf{0}, (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1})$.

Logo, uma banda assintótica de confiança de coeficiente $1 - \alpha$ para o preditor linear $\mathbf{z}\beta$, $\forall \mathbf{z} \in \mathbb{R}^p$, fica dada por

$$\mathbf{z}'\hat{\beta} \pm \sqrt{c_\alpha} [\mathbf{z}'(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{z}]^{1/2}, \quad \forall \mathbf{z} \in \mathbb{R}^p,$$

em que c_α é tal que $\mathbb{P}(\chi_p^2 \leq c_\alpha) = 1 - \alpha$.

Aplicando a transformação logito podemos, equivalentemente, encontrar uma banda de confiança de coeficiente $1 - \alpha$ para $\pi(\mathbf{z})$, dada por

$$\frac{\exp \left[\mathbf{z}' \hat{\boldsymbol{\beta}} \pm \sqrt{c_{\alpha}} \left(\mathbf{z}' (\mathbf{X}' \mathbf{W} \mathbf{X})^{-1} \mathbf{z} \right)^{1/2} \right]}{1 + \exp \left[\mathbf{z}' \hat{\boldsymbol{\beta}} \pm \sqrt{c_{\alpha}} \left(\mathbf{z}' (\mathbf{X}' \mathbf{W} \mathbf{X})^{-1} \mathbf{z} \right)^{1/2} \right]}, \quad \forall \mathbf{z} \in \mathbb{R}^p.$$

É importante observarmos que \mathbf{z} é um vetor $p \times 1$ que varia livremente no \mathbb{R}^p , enquanto \mathbf{X} é uma matriz fixa com os valores das variáveis explicativas.

Probito e complemento log-log

Modelo probito

Para o modelo probito temos que

$$\pi_i = \mu_i = F(\eta_i),$$

onde F é a f.d.a da distribuição normal padrão.

O MLG com função de ligação de probito, é dado por

$$\Phi^{-1}(\pi_i) = \eta_i = \sum_j \beta_j x_{ij}$$

ou

$$\pi_i = \Phi \left(\sum_j \beta_j x_{ij} \right).$$

Como interpretamos a magnitude de β_j ?

Temos que, para o modelo probito, a taxa de mudança em π_i quando a preditora j muda, fixando as outras covariáveis preditoras, é dada por

$$\frac{\partial \pi_i}{\partial x_{ij}} = \beta_j \phi \left(\sum_{j=1}^p \beta_j x_{ij} \right),$$

onde $\phi(\cdot)$ é a f.d.p da normal padrão.

Essa taxa é máxima (e igual a $0,40\beta_j$) quando $\sum_{j=1}^p \beta_j x_{ij} = 0$ para o qual $\pi_i = 1/2$.

Com função da preditora j , a curva de resposta probito para π_i (ou para $1 - \pi_i$, quando $\beta_j < 0$) tem a forma da f.d.a de uma normal com desvio padrão $\frac{1}{|\beta_j|}$.

É possível motivar o modelo probito como um modelo de variável latente. Suponha que exista uma variável aleatória auxiliar

$$Y^* = X'\beta + \varepsilon,$$

onde $\varepsilon \sim N(0, 1)$. Então Y pode ser visto como um indicador para saber se essa variável latente é positiva

$$Y = \begin{cases} 1, & \text{se } Y^* > 0 \\ 0, & \text{caso contrário} \end{cases} = \begin{cases} 1, & \text{se } -\varepsilon < X'\beta, \\ 0, & \text{caso contrário.} \end{cases}$$

O uso da distribuição normal padrão não causa perda de generalidade em comparação com o uso de uma distribuição normal com média e desvio padrão arbitrários, pois a adição de uma quantidade fixa à média pode ser compensada subtraindo a mesma quantidade do intercepto e multiplicando o desvio padrão de uma quantia fixa pode ser compensado multiplicando os pesos pela mesma quantia.

Para ver que os dois modelos são equivalentes, observe que

$$\begin{aligned}\mathbb{P}(Y = 1 \mid X) &= \mathbb{P}(Y^* > 0) \\ &= \mathbb{P}(X'\beta + \varepsilon > 0) \\ &= \mathbb{P}(\varepsilon > -X'\beta) \\ &= \mathbb{P}(\varepsilon < X'\beta) && \text{por simetria da distribuição normal} \\ &= \Phi(X'\beta) = \pi_i.\end{aligned}$$

Para a estimação dos parâmetros do modelo, podemos resolver o sistema de equações de verossimilhanças usando o algoritmo de Escore de Fisher ou o algoritmo Newton-Raphson.

Ambos produzem as estimativas de MV, mas o algoritmo de Newton-Raphson fornece erros padrão ligeiramente diferentes pois inverte a matriz de informação observada para estimar a matriz de covariância, enquanto o Escore Fisher usa a matriz de informação esperada.

Como visto, esses algoritmos diferem para funções de ligações que não sejam a função canônica.

Modelo complemento log-log

Para o modelo complemento log-log temos que

$$\pi_i = \mu_i = F(\eta_i),$$

onde $F(x) = 1 - \exp(-\exp(x))$.

Então, o MLG com função de ligação complemento log-log, é dado por

$$F^{-1}(1 - \pi_i) = \log(-\log(1 - \pi_i)) = \eta_i = \sum_j \beta_j x_{ij}$$

ou

$$\pi_i = 1 - \exp \left(- \exp \left(\sum_j \beta_j x_{ij} \right) \right).$$

A função de ligação para este MLG é chamada de ligação log-log complementar ou complemento log-log, uma vez que a ligação log-log se aplica ao complemento de π_i .

Voltando ao exemplo do Desastre do Challenger

```
fit.probit <- glm(cbind(damage,6-damage) ~ temp, family=binomial(link=probit),
                 orings)
summary(fit.probit)
```

Call:

```
glm(formula = cbind(damage, 6 - damage) ~ temp, family = binomial(link = probit),
    data = orings)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.0134	-0.7761	-0.4467	-0.1581	1.9983

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	5.59145	1.71055	3.269	0.00108 **
temp	-0.10580	0.02656	-3.984	6.79e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 38.898 on 22 degrees of freedom
Residual deviance: 18.131 on 21 degrees of freedom
AIC: 34.893

Number of Fisher Scoring iterations: 6


```
fit.cloglog <- glm(cbind(damage,6-damage) ~ temp, family=binomial(link=cloglog),
                  orings)
summary(fit.cloglog)
```

Call:

```
glm(formula = cbind(damage, 6 - damage) ~ temp, family = binomial(link = cloglog),
    data = orings)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-0.9884	-0.7262	-0.4373	-0.2141	1.9520

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	10.86388	2.73668	3.970	7.20e-05 ***
temp	-0.20552	0.04561	-4.506	6.59e-06 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 38.898 on 22 degrees of freedom
Residual deviance: 16.029 on 21 degrees of freedom
AIC: 32.791

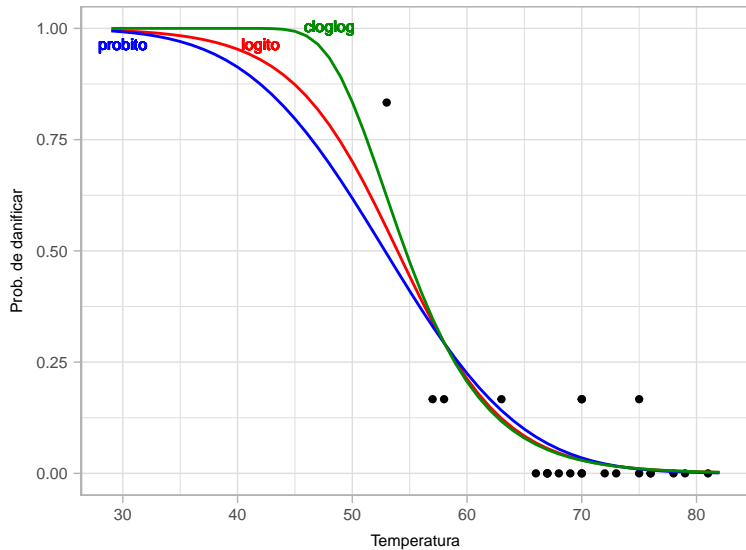
Number of Fisher Scoring iterations: 7

```
stargazer(fit.logit,fit.probit,fit.cloglog,type="text")
```

```
=====
                        Dependent variable:
-----
               cbind(damage, 6 - damage)
      logistic   probit   glm: binomial
                   link = cloglog
                (1)      (2)      (3)
-----
temp                -0.216*** -0.106*** -0.206***
                   (0.053)   (0.027)   (0.046)

Constant            11.663***  5.591***  10.864***
                   (3.296)   (1.711)   (2.737)

-----
Observations         23         23         23
Log Likelihood       -14.837   -15.447   -14.396
Akaike Inf. Crit.   33.675    34.893    32.791
=====
Note:                *p<0.1; **p<0.05; ***p<0.01
```



Valores ajustados

Valores ajustados para temperatura de 53°F.

```
beta.logit <- fit.logit$coefficients  
beta.probit <- fit.probit$coefficients  
beta.cloglog <- fit.cloglog$coefficients
```

```
ilogit(beta.logit[1]+beta.logit[2]*53)
```

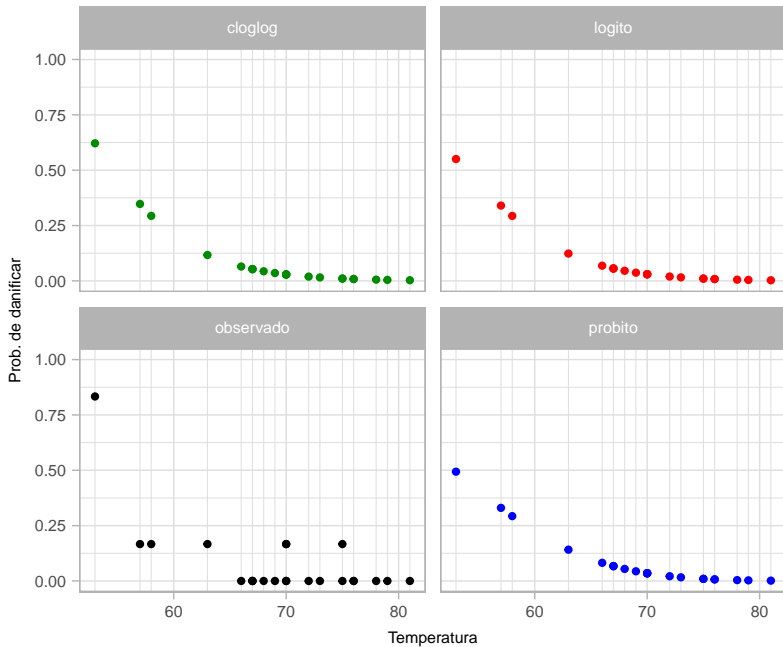
(Intercept)
0.5504788

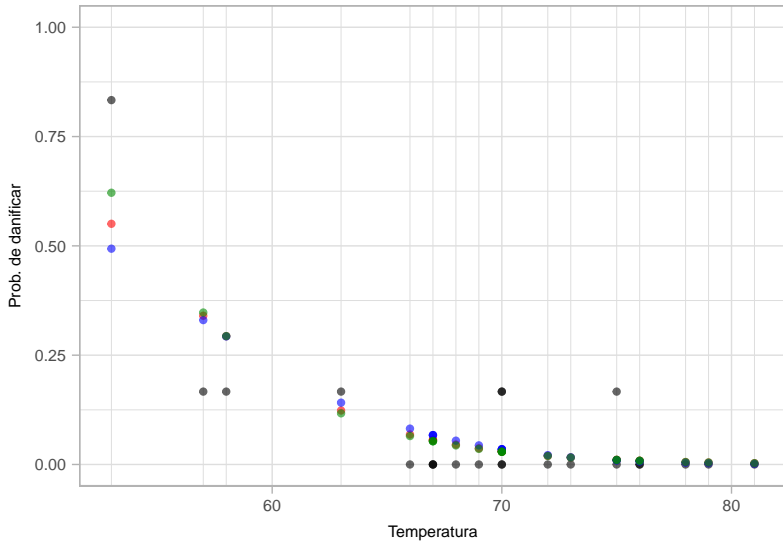
```
pnorm(beta.probit[1]+beta.probit[2]*53)
```

(Intercept)
0.4935569

```
1-exp(-exp(beta.cloglog[1]+beta.cloglog[2]*53))
```

(Intercept)
0.6214823





fc.lig cloglog logito observado probito

Interpretação dos parâmetros

Modelo probito: Como discutimos, a análise do modelo probito é baseada na f.d.a da normal padrão.

Os coeficientes do modelo probito são os efeitos nas probabilidades da f.d.a. de que a variável resposta seja igual a uma (i.e., sucesso). A seguir temos uma tabela de alguns z-escores e suas probabilidades associadas:

Probabilidade	z-escore
-2,0	0,0228
-1,0	0,1587
-0,5	0,3085
0,0	0,5000
0,5	0,6915
1,0	0,8413
2,0	0,9772

Ou seja, se vc aumenta o z-escore vc aumenta a probabilidade de sucesso.

No nosso exemplo, o intercepto é o z-escore previsto quando temos a temperatura no modelo. Podemos procurar esse z-escore e encontrar a probabilidade associada a esse z-escore.

O coeficiente da temperatura nos diz o incremento (positivo ou negativo) no z-escore para cada aumento de uma unidade na temperatura.

Assim, como $\hat{\beta}_1$ para o modelo probito é $-0,10580$, para o aumento de uma unidade na temperatura o z-escore diminui em $0,10580$.

Ou seja, para o aumento de uma unidade na temperatura nos diminuimos a probabilidade estimada de danificar o *orings*.

Obs: Não podemos comparar os coeficientes de um modelo logito e um modelo probito, podemos comparar as probabilidades preditas em algum valor de x .

Modelo complemento log-log: Agora vamos falar com mais detalhes sobre o modelo complemento log-log, considerando uma única covariável x quantitativa, temos que $\pi(x) = 1 - \exp[-\exp(\beta_0 + \beta_1 x)]$.

Para interpretá-lo, tomamos dois valores distintos da variável explicativa, x_j e x_{j+1} , e obtemos

$$\log[-\log(1 - \pi(x_{j+1}))] - \log[\log(1 - \pi(x_j))] = \beta_1(x_{j+1} - x_j),$$

de modo que

$$\frac{\log[1 - \pi(x_{j+1})]}{\log[1 - \pi(x_j)]} = \exp[\beta_1(x_{j+1} - x_j)]$$

e

$$1 - \pi(x_{j+1}) = [1 - \pi(x_j)]^{\exp[\beta_1(x_{j+1} - x_j)]}.$$

Para $x_{j+1} - x_j = 1$, a probabilidade do complemento em x_{j+1} é igual à probabilidade do complemento em x_j elevada à potência $\exp(\beta_1)$.

No nosso exemplo, para temperatura 51°F a probabilidade ajustada de não danificar (complemento) é dada por

$$1 - \pi(51) = \exp[-\exp(10,86388 - 0,2055245 * 51)] = 0,2309859$$

e a probabilidade ajustada de não danificar para temperatura 50°F é

$$1 - \pi(50) = \exp[-\exp(10,86388 - 0,2055245 * 51)] = 0,1653391.$$

A a probabilidade de não danificar na temperatura 51°F é igual a probabilidade ajustada de não danificar na temperatura 50° elevada à potência de $\exp(-0,2055245) = 0,8142202$. Por exemplo,

$$(0,1653391)^{0,8142202} \approx 0,2309859.$$

Modelo logístico: Vimos anteriormente, e é interpretado usando a razão de chances.

Predição - Matriz de confusão e Curva ROC

Predição

Quando temos uma variável resposta binária ($Y = 0$ ou 1) é necessário escolher uma regra de predição para \hat{Y} ($\hat{Y} = 0$ ou 1), uma vez que $\hat{\pi}$ está entre 0 e 1.

Uma tabela de classificação (matriz de confusão) classifica a resposta binária Y com uma predição \hat{Y} .

		Valor Observado(Y)	
		1	0
Valor Predito (\hat{Y})	1		
	0		

É intuitivo pensar que se o valor de $\hat{\pi}_i$ for grande, $\hat{Y}_i = 1$ e se $\hat{\pi}_i$ for pequeno, $\hat{Y}_i = 0$.

Mas como determinar o ponto que para os valores acima dele $\hat{Y}_i = 1$, e valores abaixo dele $\hat{Y}_i = 0$, ou seja

$$\begin{aligned}\hat{Y}_i &= 1 & \text{se } \hat{\pi}_i > \pi_0 \\ \hat{Y}_i &= 0 & \text{se } \hat{\pi}_i < \pi_0.\end{aligned}$$

O ponto π_0 é conhecido como ponto de corte.

Para um ponto de corte específico, resumos do poder preditivo da tabela de classificação são estimativas de

$$\text{sensibilidade} = \mathbb{P}(\hat{Y} = 1 | Y = 1)$$

e

$$\text{especificidade} = \mathbb{P}(\hat{Y} = 0 | Y = 0).$$

Uma desvantagem da tabela de classificação é que suas caselas dependem fortemente do ponto de corte π_0 para as predições.

Uma abordagem mais informativa considera a sensibilidade e a especificidade estimadas para todos valores possíveis de π_0 .

A sensibilidade é a taxa de verdadeiros positivos, e $\mathbb{P}(\hat{Y} = 1|Y = 0)$ (1 – especificidade) é a taxa de falsos positivos.

Um gráfico da taxa de verdadeiros positivos (sensibilidade) em função da taxa de falsos positivos (1-especificidade) quando π_0 diminui de 1 para 0 é chamado de curva ROC (Receiver Operating Characteristic Curve).

A curva ROC é uma forma bastante utilizada para determinar o ponto de corte.

Então, considerando a seguinte tabela:

		Valor Observado(Y)	
		1	0
Valor Predito (\hat{Y})	1	VP (verdadeiro positivo)	FP (falso positivo)
	0	FN (falso negativo)	VN (verdadeiro negativo)

Sensibilidade: É a proporção de verdadeiros positivos, ou seja, avalia a capacidade do modelo classificar $\hat{Y} = 1$ dado que ele é $Y = 1$:

$$\mathbb{P}(\hat{Y} = 1|Y = 1) = VP/(VP + FN).$$

Especificidade: É a proporção de verdadeiros negativos, isto é, avalia a capacidade do modelo predizer $\hat{Y} = 0$ dado que ele realmente é $Y = 0$:

$$\mathbb{P}(\hat{Y} = 0|Y = 0) = VN/(VN + FP).$$

Verdadeiro Preditivo Positivo: $\mathbb{P}(Y = 1|\hat{Y} = 1) = VP/(VP + FP)$.

Verdadeiro Preditivo Negativo: $\mathbb{P}(Y = 0|\hat{Y} = 0) = VN/(VN + FN)$.

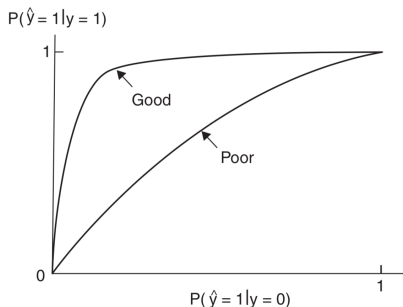


Figure 5.2 ROC curves for a binary GLM having good predictive power and for a binary GLM having poor predictive power.

Figura ilustrativa do livro texto (Agresti, A. (2015)).

Pela análise da curva ROC, escolhemos o ponto de corte referente a combinação da sensibilidade e de 1-especificidade que mais se aproxima do canto superior esquerdo do gráfico.

Desempenho do modelo de regressão logística

Para avaliar o desempenho de um modelo de regressão logística, devemos considerar poucas métricas.

- 1 AIC (Critérios de informação de Akaike):** A métrica análoga do R^2 ajustado na regressão logística é o AIC. O AIC é a medida de ajuste que penaliza o modelo pelo número de coeficientes do modelo. Portanto, sempre preferimos modelos com valor mínimo do AIC.
- 2 Desvio nulo e desvio residual:** O desvio nulo indica a resposta prevista por um modelo com nada além de um intercepto. Quanto menor o valor, melhor o modelo. O desvio residual indica a resposta prevista por um modelo na adição de variáveis independentes. Quanto menor o valor, melhor o modelo.

- 3 Matriz de Confusão:** nada mais é do que uma representação tabular dos valores reais versus os preditos. Isso nos ajuda a encontrar a precisão do modelo e a evitar ajustes excessivos.
- 4 Curva ROC:** A curva ROC resume o desempenho do modelo avaliando as compensações entre a taxa de verdadeiro positivo (sensibilidade) e a taxa de falso positivo (1- especificidade). A curva ROC resume o poder preditivo para todos os possíveis valores de $\pi_0 > 0,5$. A área sob curva (AUC), referida como índice de precisão (A) ou índice de concordância, é uma métrica de desempenho para a curva ROC. Quanto maior a área, melhor o poder de previsão do modelo. A curva ROC de um modelo preditivo perfeito tem VP igual a 1 e FP igual a 0.

- Notas de aula do Prof. Gilberto de Paula.
- Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models*. Wiley series in probability and statistics.
- Faraway, J. J. (2006). *Extending the Linear Model with R. Generalized Linear, Mixed Effects and Nonparametric Regression Models*. Chapman and Hall/CRC.
- Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.2.
<https://CRAN.R-project.org/package=stargazer>