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 distibuição binomial é a *log-odds* (*log-chances*), ou seja, o *logito*,

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

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

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:

$$\text{sucesso } (Y = 1) \text{ ou fracasso } (Y = 0).$$

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

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

e f.d.a. dada por

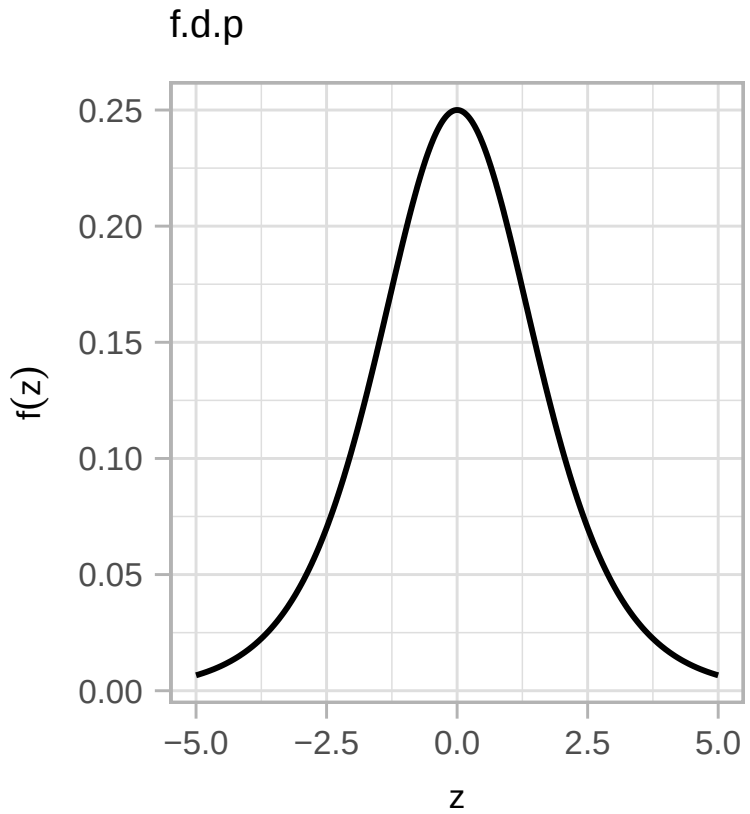
$$F(\mathrm{y}) = \frac{1}{1+\exp\left(-\frac{\mathrm{y}-\mu}{\sigma}\right)}, \quad \mathrm{y}, \mu \in \mathbb{R},\ \sigma > 0.$$

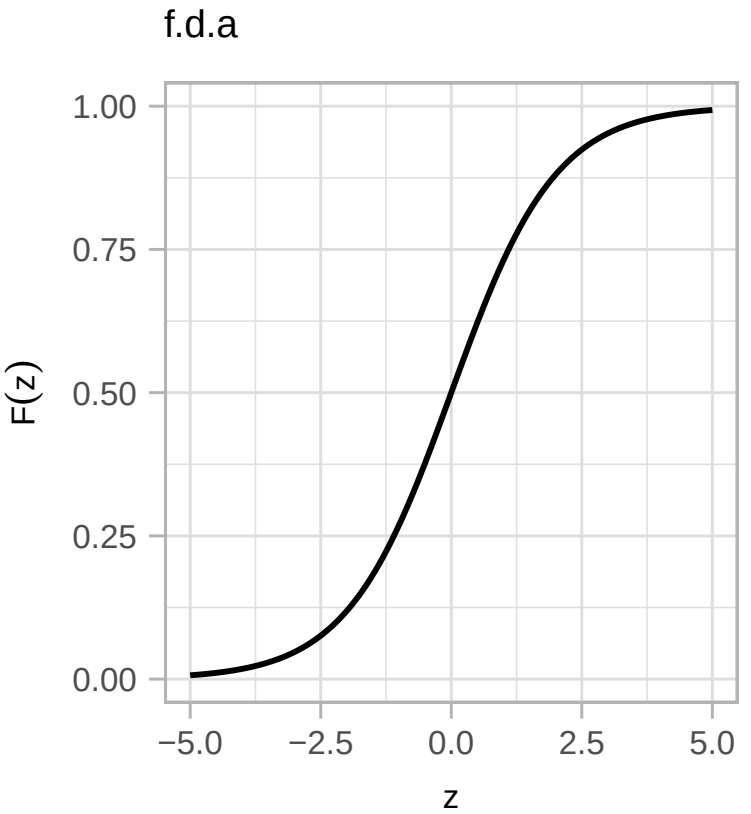
Notação: $Y \sim \text{Logística}\,(\mu, \sigma)$.

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 $\boldsymbol{X}$.

Propriedades e Interpretações

Assumindo que $Y_1, \cdots, Y_n$ são proporções binomiais independentes, com $n_i Y_i \sim Bin(n_i, \pi_i)$.

O modelo para regressão logística possui duas formulações.

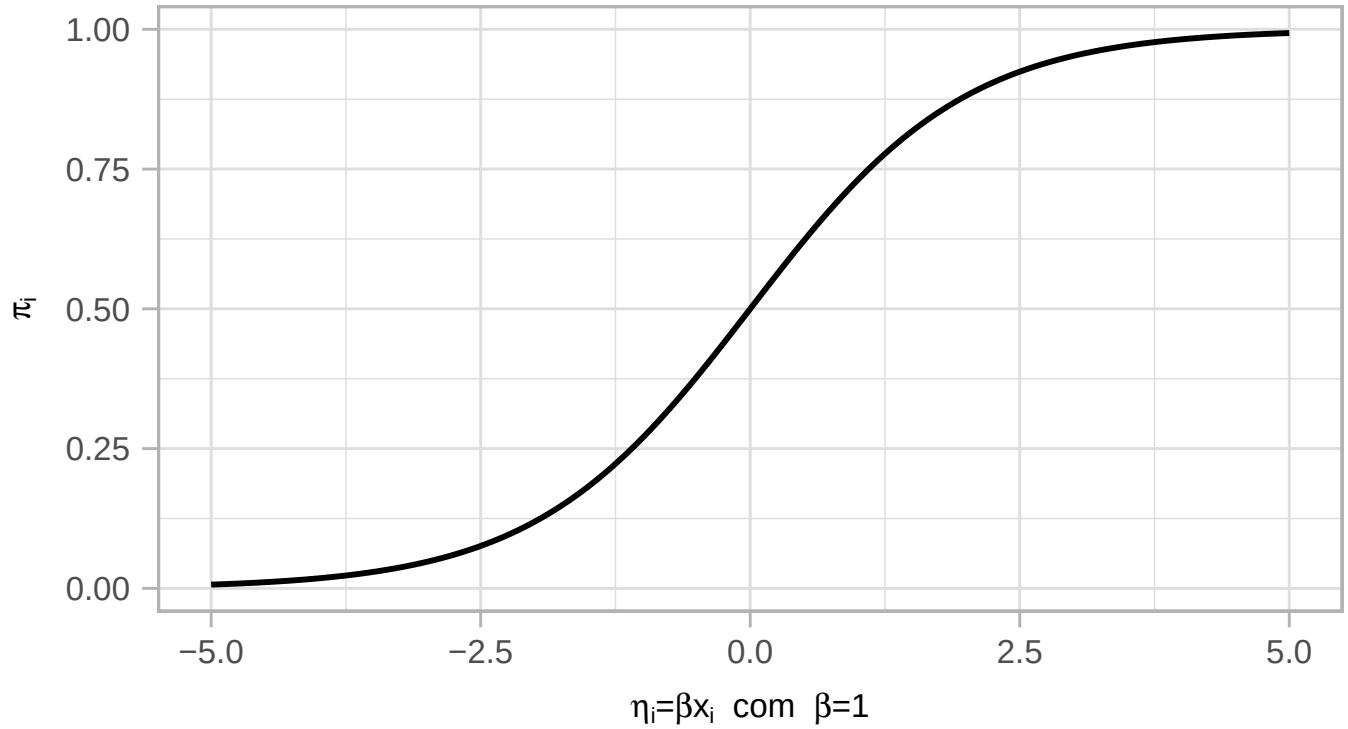
Modelo de regressão logística:

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

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

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 para π_i tem a forma da pdf 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

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

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),

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

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

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

Temos ainda que,

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

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

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

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

Exemplo Livro Faraway: Desastre do Challenger

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

Uma investigação foi iniciada para decobrir a causa do acidente; e a atenção concentrou-se nas vedações dos aneis 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^oF.

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
```

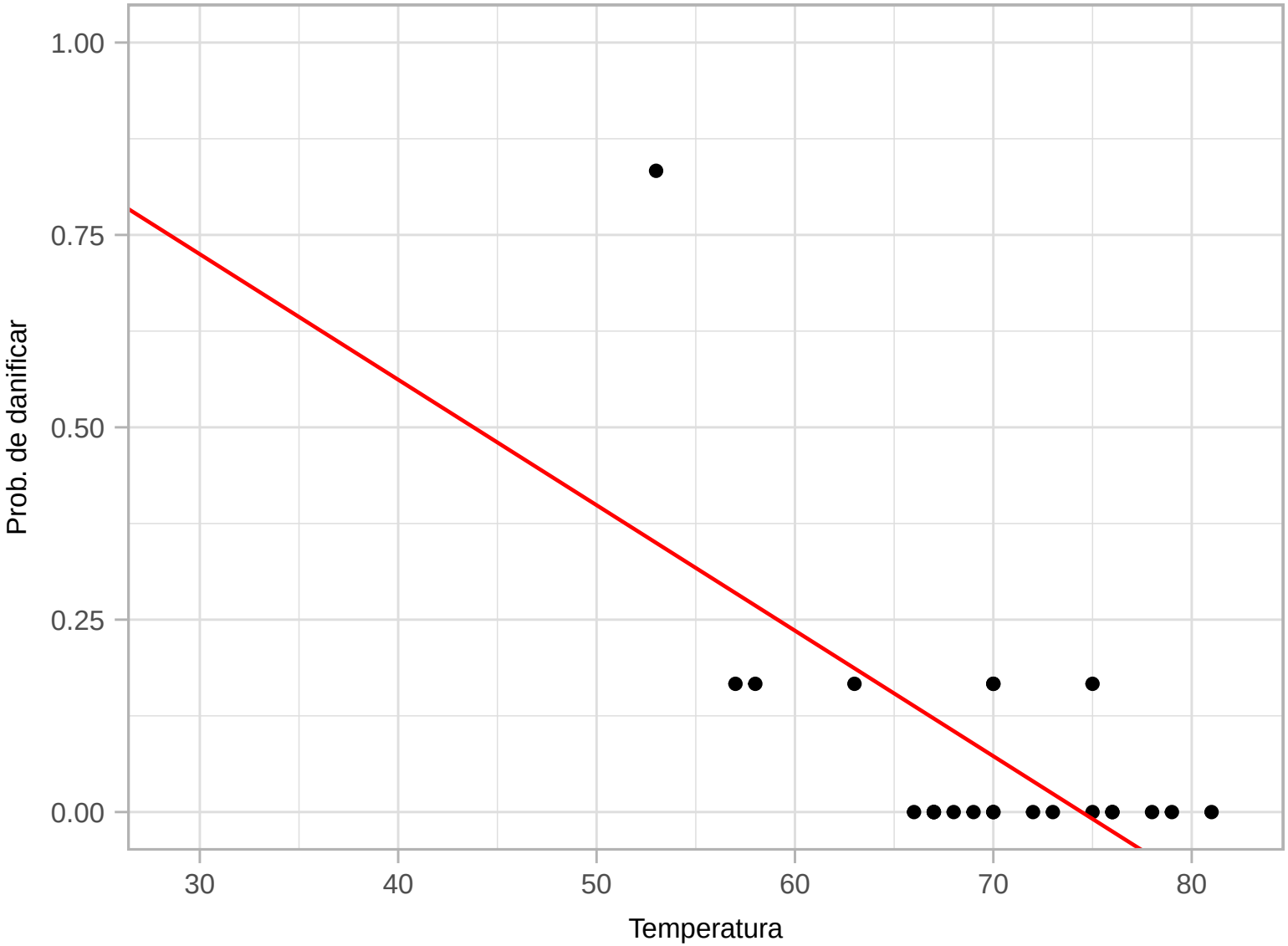
Resumindo os dados

```
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
```


1.00
0.75
0.50
0.25
0.00
Prob. de danificar
30
40
50
60
70
80
Temperatura

```r
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

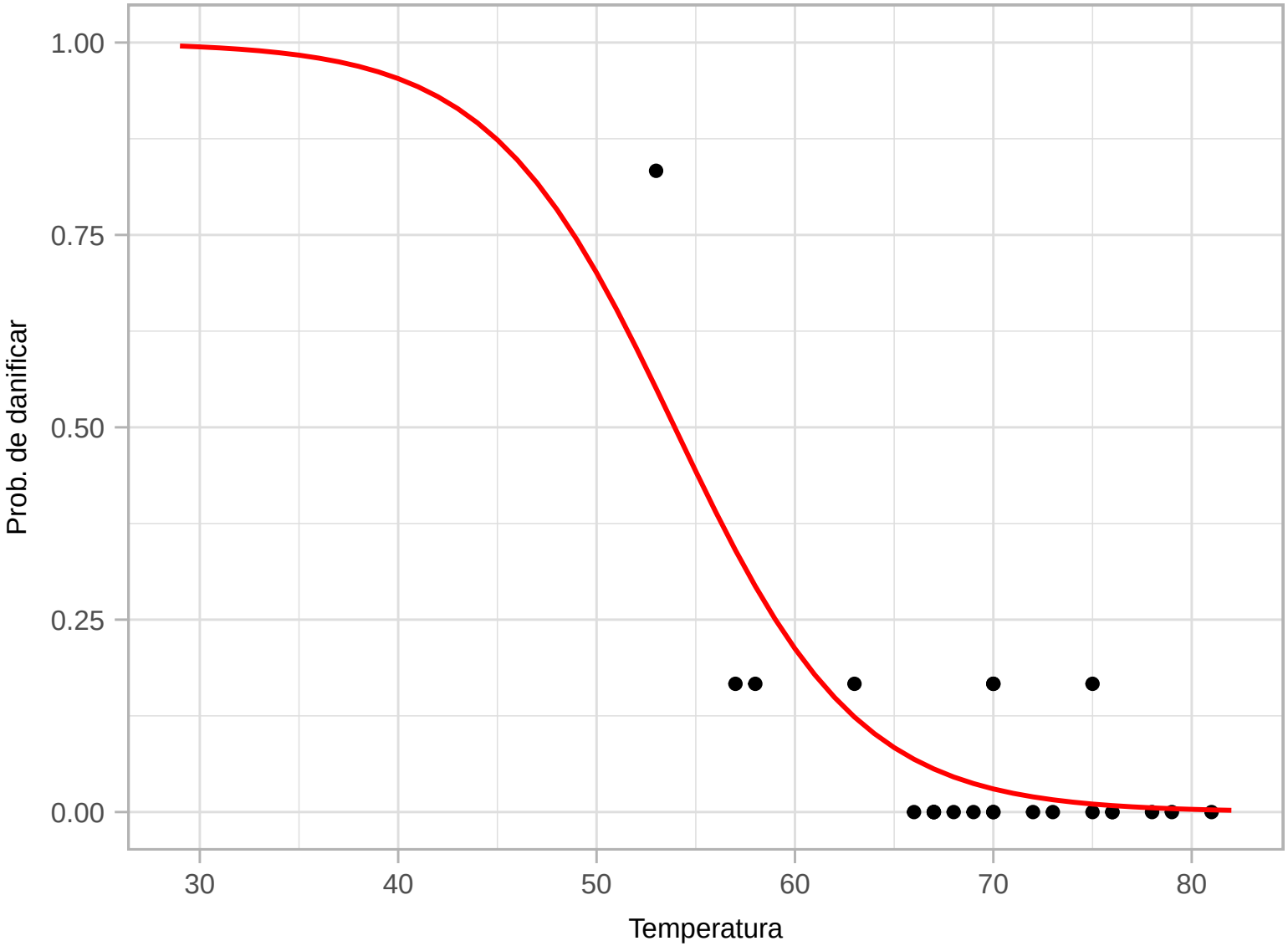
$$RC(\widehat{\beta_1}) = e^{-0,21623} = 0,80555.$$

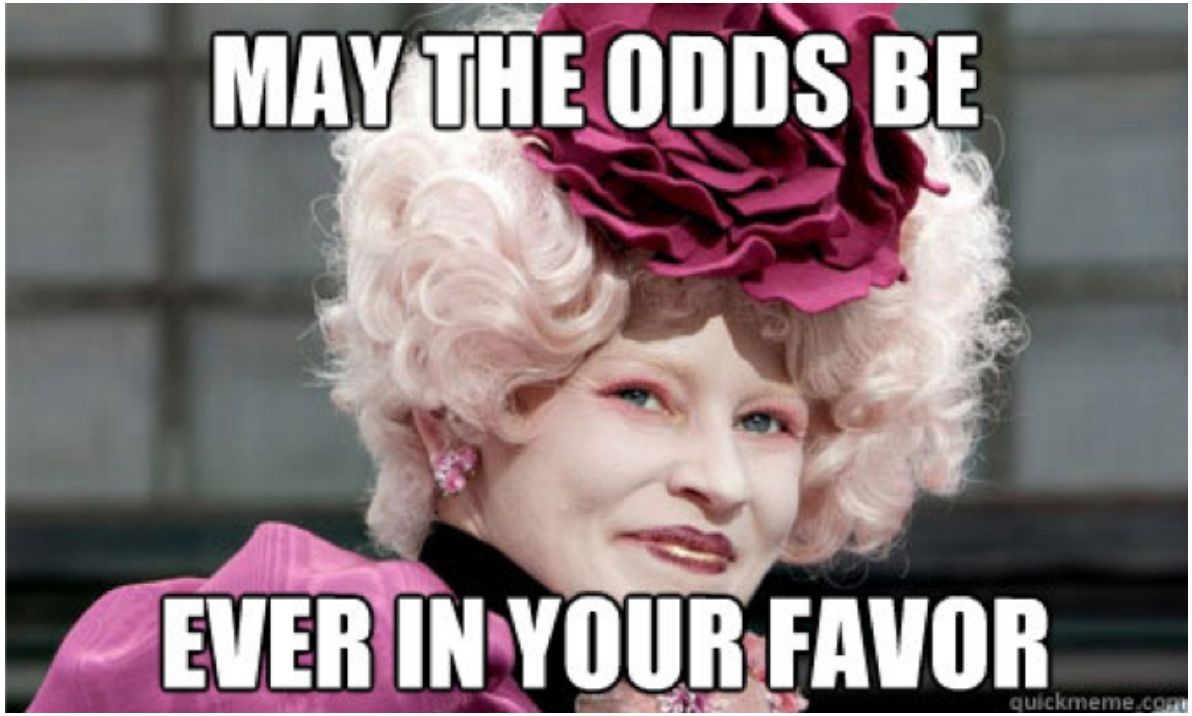
Como a RC é menor que 1, a probabilidade de danificar o *O-rings* tende a diminuir quando aumenta a temperatura do lançamento.

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

```r
cbind(orings$temp,fit.logit$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
```


1.00
0.75
0.50
0.25
0.00
Prob. de danificar
30
40
50
60
70
80
Temperatura


MAY THE ODDS BE
EVER IN YOUR FAVOR
quickmeme.com

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:

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

Portanto, temos que

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

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

Referência

- 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*. Chapaman and Hall/CRC.