MAE 5905: Introdução à Ciência de Dados Gabarito Lista 1

Alex Monito Nhancololo

2026-04-20

Critérios de avaliação

Independentemente do software empregado, aplicam-se os seguintes critérios de avaliação:

- Justificar sempre as escolhas metodológicas, não bastando apresentar apenas o código e/ou sua saída.
- Manter o documento limpo, sem # desnecessários, warnings, erros ou descrições triviais.
- Incluir tabelas e gráficos bem formatados, com títulos, legendas e rótulos claros para interpretação autônoma.
- Organizar todo o código (caso haja) em apêndice/anexo , mantendo coerência, consistência e redação técnica objetiva. Os códigos devem constar no trabalho, mas como apêndices, e não no corpo do texto.
- Se feito em grupo, deve constar a contribuição de cada membro.
- Não pode haver evidências de cópia de colegas deste curso, de anos anteriores, ChatGPT, etc.
- Excluindo os códigos, o trabalho não pode exceder 20 páginas, sendo necessário incluir apenas o essencial. Não existe número mínimo, quanto menor melhor.
- Comentários sobre funções básicas (como `set.seed()` ou outras funções triviais) não devem ser incluídos.

Dica

- O capítulo 1, seção 2.3 dasnotas de aula (link) mostra como as tabelas devem ser formatadas, seja em `Word, Overleaf/LaTeX, Markdown ou Quarto`.
- A seção 1.6 do mesmo material mostra como deve ser a saída do modelo caso haja ajuste; não é permitido apenas copiar o `summary` não formatado e colocá-lo no trabalho.
- Use `echo = TRUE, message = FALSE, warning = FALSE, comment = ""` para mostrar código e saída, ocultar mensagens automáticas e warnings, e remover o símbolo padrão (##) antes da saída do código.

Exercício 1

Num conjunto de dados, o primeiro quartil é 10, a mediana é 15 e o terceiro quartil é 20.

Seja $X = \{2.1, 7.2, 10, 11.1, 14, 16, 19, 20, 21.1, 23\}$ o conjunto de dados para uma variável aleatória contínua, ou $Y = \{2, 7, 10, 11, 14, 16, 19, 20, 21, 23\}$ para uma variável aleatória discreta.

> **Dica**
>
> Utilize a Fórmula 3.5 de Morettin e Singer (2023, p. 48). O software R aproxima os números para inteiros e aplica a versão discreta da Fórmula 3.5.

(a) A distância interquartil é 5.

Resp: **FALSO**

A distância interquartil é calculada pela fórmula:

$d_Q = Q_3 - Q_1 = 20 - 10 = 10$ (Morettin e Singer, 2023, p. 50, ponto 3.12)

(b) O valor 32 seria considerado outlier segundo o critério utilizado na construção do boxplot.

Resp: **FALSO**

Por simplicidade um valor é considerado outlier se não pertence ao intervalo $[Q_1 - 1.5 \times d_Q ; Q_3 + 1.5 \times d_Q]$:

$$
\begin{aligned}
32 &\overset{?}{\in} [Q_1 - 1.5 \times d_Q \; ; \; Q_3 + 1.5 \times d_Q] \\
32 &\overset{?}{\in} [10 - 1.5 \times 10 \; ; \; 20 + 1.5 \times 10] \\
32 &\in [5 \; ; 35]
\end{aligned}
$$

Assim, o valor 32 não é um outlier, pois está dentro dos limites inferior ($Q_1 - 1.5 \times d_Q = -5$) e superior ($Q_3 + 1.5 \times d_Q = 35$), que são padrão em várias literaturas (Morettin & Singer, 2023, p. 55, Figura 3.16 e diversas outras literaturas).

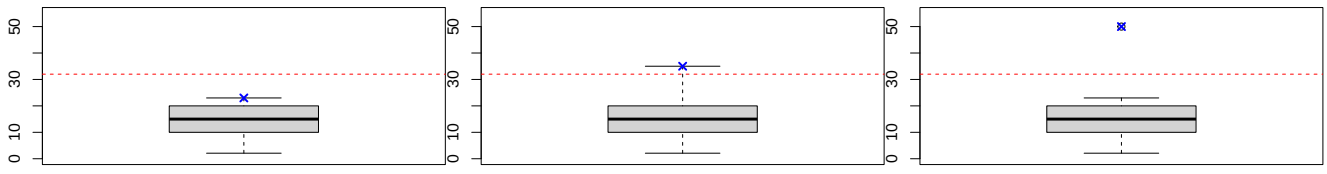
Possíveis respostas aceitáveis mediante justificativa

(1) O critério utilizado na construção do boxplot considera X_i um outlier se $X_i > \min[x_{(n)}, Q_3 + 1.5 \times d_Q]$ ou $X_i < \max[x_{(1)}, Q_1 - 1.5 \times d_Q]$ onde $x_{(1)}$ e $x_{(n)}$ são, respectivamente, o valor mínimo e máximo do conjunto de dados (Morettin & Singer, 2023, p. 54, ponto 3.4, e R/RStudio, objeto *out*, do comando `boxplot.stats` na saída abaixo, onde $d_Q = \text{iqr}$ e $\text{coef} = 1.5$).

$Q_1 - 1.5 \times d_Q = 10 - 1.5 \times 10 = -5$, $\lim_{\text{inf}} = \max[x_{(1)}, Q_1 - 1.5 \times d_Q] = \max[x_{(1)}, -5] > 32$?, considerando a descrição do exercício $(Q_1, Q_2 = \text{md}, Q_3)$, **não**.

$Q_3 + 1.5 \times d_Q = 20 + 1.5 \times 10 = 35$, $\lim_{\text{sup}} = \min[x_{(n)}, Q_3 + 1.5 \times d_Q] = \min[x_{(n)}, 35] < 32$? **Depende**.

Como o primeiro quartil é 10, temos certeza de que, se 32 for um outlier, será um outlier superior. Usando o critério (1) para a construção do boxplot, para decidir se 32 é outlier ou não, **vai depender do valor de** $x_{(n)}$.



(a) $X_{(n)} = 23$ (b) $X_{(n)} = 35$ (c) $X_{(n)} = 50$

Figura 1: Comparação de Boxplots com diferentes valores extremos ($X_{(n)}$)

Tabela 1: Análise de quartis e limites de outliers

Estatística	Conjunto X	Conjunto X1	Conjunto X2
Min.	2.10	2.10	2.10
1st Qu.	10.28	10.28	10.28
Median	15.00	15.00	15.00
Mean	14.35	15.74	17.24
3rd Qu.	19.75	19.75	19.75
Max.	23.00	35.00	50.00
Limite Superior (1.5*IQR)	33.96	33.96	33.96

A Figura 1 mostra que, em caso de presença de outliers nos dados, o R ajusta os *whiskers* (bigodes) para o n-ésimo valor mais próximo so conjunto dados que não é outlier (ver Tabela 1).

(c) A mediana ficaria alterada de 2 unidades se um ponto com valor acima do terceiro quartil fosse substituído por outro 2 vezes maior.

Resp: **FALSO**

Tabela 2: Resumo de Y

Min.	X1st.Qu.	Median	Mean	X3rd.Qu.	Max.
2	10.25	15	14.3	19.75	23

Só lembrando!!

A mediana med(X) de um conjunto ordenado de n elementos $x_1, x_2, \ldots, x_n$ pode ser definida por:

$$md(x) = \begin{cases} x_{(\frac{n+1}{2})}, \text{ se } n \text{ for ímpar} \\ \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}, \text{se } n \text{ for par} \end{cases}$$

onde x_i representa os valores ordenados do conjunto (ver Morettin e Singer, 2023, p. 47, ponto 3.3) .

Justificativa

- (1) O exercício fala sobre substituir um ponto e não aumentar a quantidade de pontos.
- (2) A mediana não é influenciada por valores extremos (ver Figura 1). Nas Figura 1a, Figura 1b e Figura 1c a substituição do 23 por 35 e 50 não alterou a mediana.

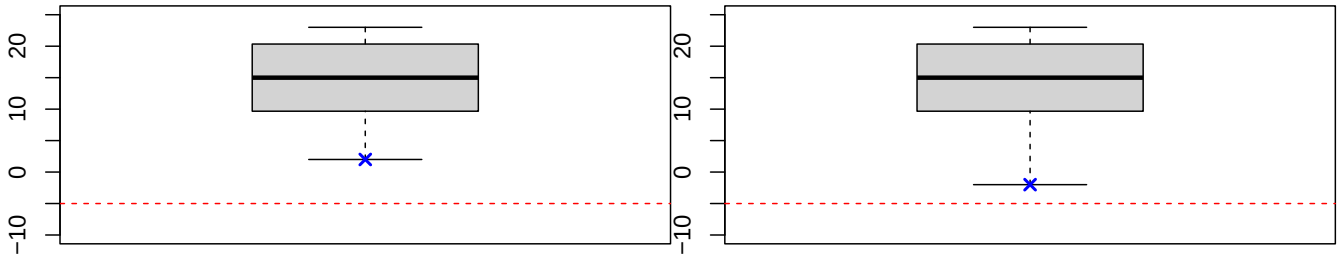
- (3) O valor a ser retirado não é um ponto de massa (ponto que divide os dados ao meio) e muito menos próximo a este, o que implica que a mediana não seria alterada.

(d) O valor mínimo é maior do que zero.

Resp: **FALSO**

Tabela 3: Efeito da alteração do valor mínimo

Estatistica	X_j	X_k
Mínimo	2.00	-2.00
1º Quartil (Q1)	9.67	9.67
Mediana	15.00	15.00
Média	14.30	13.90
3º Quartil (Q3)	20.34	20.34
Máximo	23.00	23.00
IQR	10.00	10.00
Limite Inf. (Q1 - 1.5*IQR)	-5.00	-5.00



(a) Dados com $X_{(1)} = 2$ (b) Dados com $X_{(1)} = -2$

Figura 2: Efeito do valor mínimo no primeiro quartil

Como pode ser visto na Tabela 3 e na Figura 2 , as duas listas de dados satisfazem o descrito no exercício, mas o valor mínimo não é necessariamente maior que zero .

sugestão de leitura

Mazarei, A., Sousa, R., Mendes-Moreira, J., Molchanov, S., & Ferreira, H. M. (2025). **Online boxplot derived outlier detection.** International journal of data science and analytics, 19(1), 83-97.

Exercício 2

Tabela 4: Distribuição de frequência do número de vasos

Número de vasos	Frequência absoluta (fi)	Frequência acumulada (%)
0 - 5	8 (12%)	12%
5 - 10	23 (35%)	47%
10 - 15	12 (18%)	65%
15 - 20	9 (14%)	79%
20 - 25	8 (12%)	91%
25 - 30	6 (9%)	100%

Resp: **Alternativa correta B**

Justificativa: A mediana é definida como o valor que divide o conjunto de dados em duas partes de igual frequência acumulada, correspondendo ao ponto de 50%; ao analisarmos a Tabela 4, observamos que a classe $5 - 10$ acumula 47% dos dados, enquanto a classe subsequente, $10 - 15$, atinge um acumulado de 65%, o que indica que o marco de 50% é alcançado dentro deste segundo intervalo, confirmando que a mediana pertence à classe $10 - 15$.

Fórmula caso seja de interesse

Para calcular quantis (quartis, decis e percentis) em uma distribuição de frequências, pode-se utilizar a fórmula:

$Q(p) = Lim_{inf} + \left(\frac{p \times n - F}{f}\right) \times h$

onde Q(p) é quantil desejado. p é Posição relativa/ordem do quantil. Lim_{inf} é o limite inferior da classe onde o quantil se encontra. n é número total de observações (soma das frequências absolutas). F é afrequência acumulada da classe anterior à classe do quantil. f é frequência absoluta da classe do quantil e h é amplitude da classe (diferença entre os limites inferior e superior de um intervalo).

Exercício 3

Tabela 5: Estatísticas descritivas do VO2MAX por grupo

Grupo	n	Média	Mediana	Desvio Padrão
Normais	56	1845	1707	795
Cardiopatas	57	1065	984	434
DPOC	46	889	820	381

Tabela 6: Estatísticas descritivas do VCO2MAX por grupo

Grupo	n	Média	Mediana	Desvio Padrão
Normais	56	2020	1847	918
Cardiopatas	57	1206	1081	479
DPOC	46	934	860	430

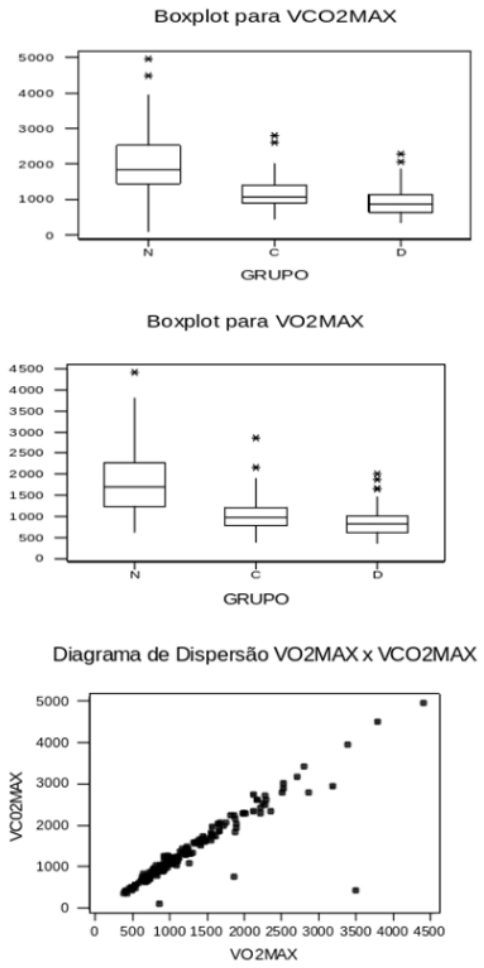


Figura 3: Gráficos para o Exercício 3

Resp: Considerando ambos os critérios, o grupo Normais apresenta a maior variabilidade global. Além de possuir um dos maiores coeficientes de variação, demonstra a maior amplitude de dispersão e sensibilidade a valores extremos nas distribuições de VO2 e VCO2.

Justificativa

Como os grupos apresentam médias distintas (conforme as Tabela 5 e Tabela 6), a comparação direta pelo desvio padrão (s) pode ser enganosa. Assim, a variabilidade deve ser analisada de forma relativa pelo coeficiente de variação (CV) ou pela distância interquartílica (IQR), observada na Figura 3.

$$CV = \left(\frac{s}{\bar{x}}\right) \times 100$$

1. Análise pelo Coeficiente de Variação

- Grupo VO2MAX: O grupo Normais apresenta a maior variabilidade relativa ($CV_N = 795 \cdot 100/1845 = 43,09\%$), seguido de perto pelo grupo DPOC ($CV_{DPOC} = 42,86\%$) e pelos Cardiopatas ($CV_C = 40,75\%$).
- Grupo VCO2MAX: O grupo DPOC possui a maior variabilidade relativa ($CV_{DPOC} = 46,04\%$), superando levemente os Normais ($CV_N = 45,45\%$) e os Cardiopatas ($CV_C = 39,72\%$).

2. Análise pela Figura 3

Embora o CV aponte o grupo DPOC com maior variabilidade relativa no VCO2MAX, a análise visual da Figura 3 mostra que o grupo Normais (N) detém a maior dispersão absoluta dos dados porque:

1. Caixas mais "altas" (maior amplitude interquartílica);
2. *Whiskers* (hastes) mais extensos;
3. Presença de *outliers* em patamares muito superiores aos demais grupos.

b) Como mostram as Tabela 5 e Tabela 6, os valores de consumo de oxigênio e CO_2 diminuem conforme passamos do grupo saudável para o de cardiopatas e, por fim, para o grupo DPOC. Em todos os cenários, tanto a média quanto a mediana seguem essa mesma queda de desempenho. Além disso, a média é sempre um pouco maior que a mediana em todos os grupos, o que acontece porque os valores muito altos observados na Figura 3 acabam puxando a média para cima.

(c) Resp:

Distância Interquartílica (IQR): Observando a Figura 3, o grupo Normais apresenta a maior IQR (caixas mais altas), indicando maior dispersão central. Já os grupos C e D possuem distâncias menores e muito semelhantes entre si.

Distribuição Normal: Não é razoável assumir a normalidade. Na Figura 3 observa-se uma presença de vários *outliers* (asteriscos) e uma assimetria à direita(bigodes superiores mais longos) e pelas médias maiores que as medianas.

(d) Os asteriscos na Figura 3 representam os *outliers* (valores atípicos ou discrepantes). Eles indicam indivíduos que apresentaram resultados muito acima do esperado para o seu respectivo grupo.

(e) O modelo mais adequado para descrever a relação entre o consumo máximo de CO_2 (y_i) e o de O_2 (x_i) é uma Regressão Linear Simples, expressa por $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, visto que o diagrama de dispersão na Figura 3 mostra uma tendência linear com forte correlação ($r = 0,92$) entre essas variáveis. No entanto, a presença de outliers (ver Figura 3) indica que os resíduos não seguem uma distribuição normal perfeita, sugerindo que, embora a função seja uma reta, pode ser necessário utilizar métodos de estimação mais robustos ou ponderados para evitar que valores extremos distorçam o ajuste do modelo.

(f) Sim, existem pontos que exigem verificação, como os *outliers* na Figura 3 que sugerem possíveis erros de digitação ou calibração. No diagrama de dispersão, destacam-se três registros abaixo da tendência linear; como o consumo de oxigênio (VO_2) e a produção de gás carbônico (VCO_2) são processos metabolicamente acoplados, é improvável consumir altas doses de O_2 sem a liberação proporcional de CO_2.

Exercício 4

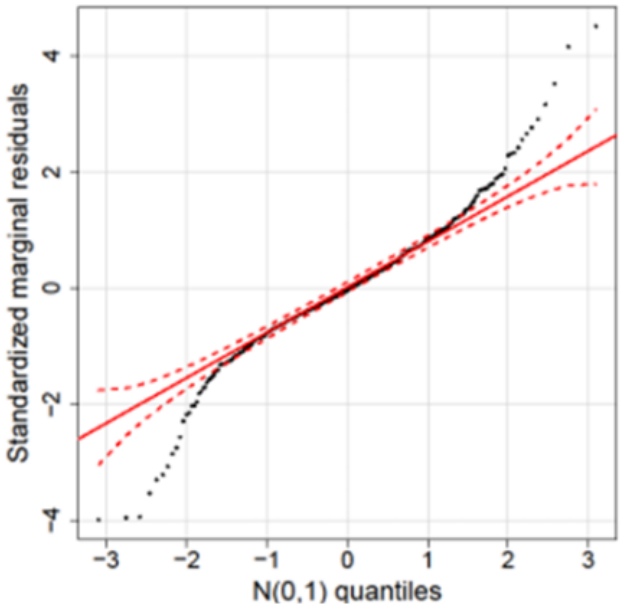


Figura 4: Gráficos para o Exercício 3

a) Alternativa correta: (D)

Na Figura 4 observa-se que os resíduos desviam-se nas caudas (extremidades), ficando fora dos limites mínimos e máximos (bandas de confiança) para quantis abaixo de -1 e acima de 1, evidenciando que os resíduos padronizados não são normalmente distribuídos. Como o desvio ocorre em ambos os extremos, entre -1 e 1, isso remete à simetria.

Exercício 5

Tabela 7: Estimativas dos parâmetros do modelo de regressão

Parâmetro	Estimativa	Erro Padrão	Valor p
α	0.690	0.120	< 0.01
β	0.330	0.100	< 0.01
γ	-0.030	0.005	< 0.01

O modelo de regressão logística para a preferência pelo refrigerante Kcola é definido por:

$$\log\left[\frac{\pi_i(x_i, w_i)}{1-\pi_i(x_i, w_i)}\right] = \alpha + \beta x_i + \gamma(w_i - 5), \tag{1}$$

onde x_i é o gênero (1 = Masculino, 0 = Feminino) e w_i a idade. A centralização $(w_i - 5)$ permite que o intercepto α represente o logito de uma menina de 5 anos. Considerando um nível de significância de 5% (correspondente a um intervalo de confiança de 95%), podemos afirmar que, como o valor-p é menor que 5% para cada parâmetro (α, β e γ), a idade (w_i) e o gênero (x_i) são variáveis importantes para determinar

a preferência pelo refrigerante (ver Tabela 7). Consequentemente, os intervalos de confiança estimados para os parâmetros α, β e γ, dados por

$$IC_{(\beta_j,\,95\%)} = \hat{\beta}_j \pm z_{\alpha/2} \cdot s.e.(\hat{\beta}_j), \tag{2}$$

não incluirão o valor zero. Ao testar esses parâmetros ($H_0 : \beta_j = 0$), estamos verificando se são estatisticamente diferentes de zero. Se o valor-p (ver Tabela 7) fosse maior que 5%, não teríamos evidências suficientes para rejeitar a hipótese nula de que os parâmetros são iguais a zero e, nesse caso, o IC incluiria o valor zero.

> **!** Importante
>
> Para a interpretação dos parâmetros, existem três formas: (1) Escala do Logito (Log-Odds), (2) Escala de Chance (Odds) e (3) Escala de Probabilidade.

- Na Escala do Logito, avaliamos a força direta da tendência. A relação é vista de forma linear e aditiva:

$$\text{Logito}(\pi) = \alpha + \beta x_i + \gamma(w_i - 5)$$

Aqui, cada unidade a mais na idade ou a mudança de gênero altera o logaritmo da chance (log-odds) da preferência de forma constante em γ ou β unidades, respectivamente. Quando temos mais de uma variável, o procedimento padrão é estudar o efeito isolado de cada uma, mantendo as demais constantes (ceteris paribus). Atribuímos zero às demais variáveis apenas quando queremos observar o efeito especificamente no nosso grupo de referência ou para interpretar o ponto de partida (intercepto) do modelo.

- Na Escala de Chance, o efeito torna-se multiplicativo. Partindo da Equação 1 e aplicando a função exponencial para remover o logaritmo, temos:

$$\begin{aligned} \frac{\pi_i}{1-\pi_i} &= \exp(\alpha + \beta x_i + \gamma(w_i - 5)) \\ \text{Chance} &= e^{\alpha} \cdot e^{\beta x_i} \cdot e^{\gamma(w_i - 5)} \end{aligned}$$

Para entender o impacto prático de um aumento unitário, comparamos o estado de um indivíduo (X) com o estado após um incremento ($X+1$), como o envelhecimento de exatamente um (1) ano. Calculamos a Razão de Chances (RC) para isolar esse fator de mudança:

$$\begin{aligned} \text{RC} &= \frac{\text{Chance}(X+1)}{\text{Chance}(X)} \\ \text{RC} &= \frac{e^{\alpha+\beta(X+1)}}{e^{\alpha+\beta X}} \\ \text{RC} &= \frac{e^{\alpha+\beta X} \cdot e^{\beta}}{e^{\alpha+\beta X}} \\ \text{RC} &= e^{\beta} \end{aligned}$$

Na prática, o valor e^{β} é o multiplicador, ele nos diz quantas vezes a chance do evento aumenta em relação ao estado anterior, independentemente de onde o indivíduo comece na escala.

- Por fim, na Escala de Probabilidade, observamos o comportamento real da probabilidade em uma curva em formato de "S". Primeiro, isolamos a probabilidade (π) a partir da Equação 1 considerando Logito $= \eta$:

$$\begin{aligned} \frac{\pi}{1-\pi} &= e^{\eta} \\ \pi &= e^{\eta}(1-\pi) \\ \pi + \pi e^{\eta} &= e^{\eta} \\ \pi(1+e^{\eta}) &= e^{\eta} \\ \pi &= \frac{e^{\eta}}{1+e^{\eta}} = \frac{1}{1+e^{-\eta}} \end{aligned}$$

Para saber quão rápido a preferência muda, usamos a derivada (efeito marginal) pela regra da cadeia:

$$\begin{aligned} \frac{\partial \pi}{\partial X_j} &= \frac{\partial \pi}{\partial \eta} \cdot \frac{\partial \eta}{\partial X_j} \\ \frac{\partial \pi}{\partial X_j} &= \left[\frac{d}{d\eta}(1+e^{-\eta})^{-1}\right] \cdot \beta_j \\ \frac{\partial \pi}{\partial X_j} &= \left[\frac{e^{-\eta}}{(1+e^{-\eta})^2}\right] \cdot \beta_j \\ \frac{\partial \pi}{\partial X_j} &= \left[\frac{1}{1+e^{-\eta}} \cdot \frac{e^{-\eta}}{1+e^{-\eta}}\right] \cdot \beta_j \\ \frac{\partial \pi}{\partial X_j} &= \beta_j \cdot \pi(1-\pi) \end{aligned} \tag{3}$$

A Equação 3 mostra que a probabilidade não sobe de forma igual para todos. Em crianças que já apresentam forte rejeição ou aceitação total (extremos), um ano a mais quase não altera a probabilidade. O impacto da mudança é maior naquelas que se encontram no ponto de máxima incerteza (ou região de indiferença, onde $\pi = 0,5$), pois é nesse estágio que o termo $\pi(1-\pi)$ atinge seu valor máximo, tornando o modelo extremamente sensível e qualquer variação na idade muito mais impactante para o resultado final.

> **!** Importante
>
> A interpretação em probabilidade é ideal por ser a escala mais intuitiva para a tomada de decisão, traduzindo coeficientes abstratos para o risco real de 0% a 100%. Além disso, ela respeita a natureza não-linear do fenômeno, mostrando que o impacto real de uma mudança depende do estado prévio do indivíduo na curva.

a) Interpretação de cada parâmetro

$\Longrightarrow$ Para α, considere uma criança do gênero feminino ($x_i = 0$) com idade $w_i = 5$ anos:

$$
\begin{aligned}
\log\left[\frac{\pi_i(0,5)}{1-\pi_i(0,5)}\right] &= \alpha + \beta\cdot 0 + \gamma(5-5)\\
\log\left[\frac{\pi_i(0,5)}{1-\pi_i(0,5)}\right] &= \alpha\\
\frac{\pi_i(0,5)}{1-\pi_i(0,5)} &= \exp(\alpha)\\
\frac{\pi_i(0,5)}{1-\pi_i(0,5)} &= \exp(0.69)\\
\frac{\pi_i(0,5)}{1-\pi_i(0,5)} &\approx 1.99\\
\pi_i(0,5) &= \frac{1,99}{1+1,99} \approx 0,666
\end{aligned}
\tag{4}
$$

- Escala do Logito: O valor $\alpha = 0,69$ representa o logaritmo da chance de preferência pelo refrigerante Kcola.
- Escala de Chance (Odds): A chance de preferência é de $1,99$, indicando que, para este perfil, a preferência pelo produto é aproximadamente o dobro da não preferência. Equivalentemente, podemos dizer que a chance de uma menina de 5 anos preferir o refrigerante é 99% superior à chance de não preferir.
- Escala de Probabilidade: A probabilidade estimada de uma menina de 5 anos preferir o refrigerante é de $66,6\%$. Em termos práticos, estima-se que duas em cada três crianças (2/3) com este perfil clínico prefiram a bebida.

$\Longrightarrow$ Para β, considere $x_i = 1$ e $x_j = 0$, mantendo a idade w fixa:

$$
\begin{aligned}
\log\left[\frac{\frac{\pi_i(1,w)}{1-\pi_i(1,w)}}{\frac{\pi_j(0,w)}{1-\pi_j(0,w)}}\right] &= \log\left[\frac{\pi_i(1,w)}{1-\pi_i(1,w)}\right] - \log\left[\frac{\pi_j(0,w)}{1-\pi_j(0,w)}\right]\\
\log\left[\frac{\frac{\pi_i(1,w)}{1-\pi_i(1,w)}}{\frac{\pi_j(0,w)}{1-\pi_j(0,w)}}\right] &= [\alpha + \beta\cdot 1 + \gamma(w-5)] - [\alpha + \beta\cdot 0 + \gamma(w-5)]\\
\log\left[\frac{\frac{\pi_i(1,w)}{1-\pi_i(1,w)}}{\frac{\pi_j(0,w)}{1-\pi_j(0,w)}}\right] &= \alpha + \beta + \gamma(w-5) - \alpha - \gamma(w-5)\\
\log\left[\frac{\frac{\pi_i(1,w)}{1-\pi_i(1,w)}}{\frac{\pi_j(0,w)}{1-\pi_j(0,w)}}\right] &= \beta \quad (*)\\
\frac{\frac{\pi_i(1,w)}{1-\pi_i(1,w)}}{\frac{\pi_j(0,w)}{1-\pi_j(0,w)}} &= \exp(\beta) = \exp(0.33) \approx 1.39 \quad (**)\\
\frac{\pi_i(1,w)}{1-\pi_i(1,w)} &= 1.39\left[\frac{\pi_j(0,w)}{1-\pi_j(0,w)}\right] \quad (***)
\end{aligned}
$$

Assim, $\beta = 0.33$ $(*)$ representa o logaritmo da razão de chances de preferir o refrigerante Kcola comparando crianças do gênero masculino com as do gênero feminino, mantendo a idade fixa. O valor $(**)$ representa a razão de chances de preferir Kcola entre crianças do gênero masculino e feminino, com idade fixa. A equação $(***)$ indica que, fixada a idade (w), a chance de uma criança do gênero masculino preferir Kcola é 1.39 vezes a chance de uma criança do gênero feminino preferir Kcola. Equivalentemente, meninos

apresentam uma chance 39% maior de preferência do que as meninas, *ceteris paribus* (mantendo tudo constante).

Pela Equação 3 sabemos que a taxa de variação é dada por $\frac{\partial \pi}{\partial X_j} = \beta_j \cdot \pi(1-\pi)$. Pela Equação 4, definimos nosso perfil base com probabilidade inicial $\pi \approx 0,666$.

Definindo o Efeito Marginal como $\Delta\pi$ (a variação na probabilidade), temos:

$$\begin{aligned}
\Delta\pi &\approx \beta \cdot \pi(1-\pi) \\
\Delta\pi &\approx 0,33 \cdot 0,666 \cdot (1-0,666) \\
\Delta\pi &\approx 0,073
\end{aligned}$$

> Para uma criança que se enquadra no perfil base (feminino, 5 anos), a mudança para o gênero masculino, mantendo a idade constante, está associada a um incremento de 7,3% na probabilidade de preferência. É fundamental notar que este ganho na probabilidade não é linear; ele depende da posição inicial do paciente na curva de risco. Em idades diferentes, onde a probabilidade de partida (π) seja outra, o impacto da mudança de gênero na probabilidade será diferente, refletindo a sensibilidade variável do modelo.

$\Longrightarrow$ Para γ, considere as idades $w_i = 6$ e $w_j = 5$ com o gênero x fixo:

$$\begin{aligned}
\log\left[\frac{\frac{\pi_i(x,6)}{1-\pi_i(x,6)}}{\frac{\pi_j(x,5)}{1-\pi_j(x,5)}}\right] &= \log\left[\frac{\pi_i(x,6)}{1-\pi_i(x,6)}\right] - \log\left[\frac{\pi_j(x,5)}{1-\pi_j(x,5)}\right] \\
\log\left[\frac{\frac{\pi_i(x,6)}{1-\pi_i(x,6)}}{\frac{\pi_j(x,5)}{1-\pi_j(x,5)}}\right] &= [\alpha + \beta x + \gamma(6-5)] - [\alpha + \beta x + \gamma(5-5)] \\
\log\left[\frac{\frac{\pi_i(x,6)}{1-\pi_i(x,6)}}{\frac{\pi_j(x,5)}{1-\pi_j(x,5)}}\right] &= \alpha + \beta x + \gamma - \alpha - \beta x \\
\log\left[\frac{\frac{\pi_i(x,6)}{1-\pi_i(x,6)}}{\frac{\pi_j(x,5)}{1-\pi_j(x,5)}}\right] &= \gamma \quad (a) \\
\frac{\frac{\pi_i(x,6)}{1-\pi_i(x,6)}}{\frac{\pi_j(x,5)}{1-\pi_j(x,5)}} &= \exp(\gamma) = \exp(-0,03) \approx 0,97 \quad (b) \\
\frac{\pi_i(x,6)}{1-\pi_i(x,6)} &= 0,97\left[\frac{\pi_j(x,5)}{1-\pi_j(x,5)}\right] \quad (c)
\end{aligned}$$

Assim, $\gamma = -0,03$ (a) representa o logaritmo da razão de chances de preferência pelo refrigerante Kcola entre crianças de 6 e 5 anos de idade, com o gênero fixado. O valor (b) representa a razão de chances de preferência entre crianças de 6 e 5 anos de idade. A equação (c) informa que a chance de uma criança de 6 anos preferir Kcola, independentemente do gênero, é $0,97$ vezes a chance de uma criança de 5 anos preferir o mesmo refrigerante. Equivalentemente, observa-se uma redução de 3% $(1 - 0,97 = 0,03)$ na chance de preferência ao adicionar um ano na idade, *ceteris paribus* (mantendo tudo constante).

Caso geral: Interpretação do parâmetro γ.

Seja a idade $w_j = w_i - 1$, com $w_i \geq 6$, e o gênero x fixo:

$$\log\left[\frac{\frac{\pi_i(x,w_i)}{1-\pi_i(x,w_i)}}{\frac{\pi_j(x,w_i-1)}{1-\pi_j(x,w_i-1)}}\right] = \log\left[\frac{\pi_i(x,w_i)}{1-\pi_i(x,w_i)}\right] - \log\left[\frac{\pi_j(x,w_i-1)}{1-\pi_j(x,w_i-1)}\right]$$

$$\log\left[\frac{\frac{\pi_i(x,w_i)}{1-\pi_i(x,w_i)}}{\frac{\pi_j(x,w_i-1)}{1-\pi_j(x,w_i-1)}}\right] = [\alpha+\beta x+\gamma(w_i-5)]-[\alpha+\beta x+\gamma(w_i-1-5)]$$

$$\log\left[\frac{\frac{\pi_i(x,w_i)}{1-\pi_i(x,w_i)}}{\frac{\pi_j(x,w_i-1)}{1-\pi_j(x,w_i-1)}}\right] = \alpha+\beta x+\gamma(w_i-5)-\alpha-\beta x-\gamma(w_i-6)$$

$$\log\left[\frac{\frac{\pi_i(x,w_i)}{1-\pi_i(x,w_i)}}{\frac{\pi_j(x,w_i-1)}{1-\pi_j(x,w_i-1)}}\right] = \gamma(w_i-5-w_i+6)$$

$$\log\left[\frac{\frac{\pi_i(x,w_i)}{1-\pi_i(x,w_i)}}{\frac{\pi_j(x,w_i-1)}{1-\pi_j(x,w_i-1)}}\right] = \gamma(1)=\gamma=-0,03 \quad (*)$$

$$\frac{\frac{\pi_i(x,w_i)}{1-\pi_i(x,w_i)}}{\frac{\pi_j(x,w_i-1)}{1-\pi_j(x,w_i-1)}} = \exp(\gamma)=\exp(-0,03)\approx 0,97 \quad (**)$$

$$\frac{\pi_i(x,w_i)}{1-\pi_i(x,w_i)} = 0,97\left[\frac{\pi_j(x,w_i-1)}{1-\pi_j(x,w_i-1)}\right] \quad (***)$$

Assim, $\gamma=-0,03$ $(*)$ representa o logaritmo da razão de chances de preferência por Kcola entre duas crianças do mesmo gênero, mas que diferem na idade em apenas um ano, sendo a primeira mais velha e a segunda mais nova. O valor $(**)$ representa a razão de chances correspondente. A equação $(***)$ informa que a chance de uma criança um ano mais velha preferir Kcola, independentemente do gênero, é $0,97$ vezes a chance da criança um ano mais nova. Equivalentemente, a cada ano a mais de idade, a criança apresenta uma chance 3% menor de preferência pelo refrigerante, *ceteris paribus* (mantendo tudo constante).

Pela Equação 3 sabemos que a taxa de variação é dada por $\frac{\partial\pi}{\partial X_j}=\gamma\cdot\pi(1-\pi)$. Pela Equação 4, definimos nosso perfil base com probabilidade inicial $\pi\approx 0,666$.

Definindo o Efeito Marginal como $\Delta\pi$ (a variação na probabilidade), temos:

$$\begin{aligned}\Delta\pi &\approx \gamma\cdot\pi(1-\pi)\\ \Delta\pi &\approx -0,03\cdot 0,666\cdot(1-0,666)\\ \Delta\pi &\approx -0,007\end{aligned}$$

Assim, para uma criança que se enquadra no perfil base (feminino, 5 anos), o incremento de um ano na idade, mantendo o gênero constante, está associado a uma redução de aproximadamente $0,7\%$ na probabilidade de preferência.

> **Aviso**
>
> Dizer que a razão de chances entre duas crianças se altera em $\exp(\gamma)$ a cada ano, fixado o gênero, remete à ideia de uma relação multiplicativa constante. Ou seja, se a diferença de idade entre duas crianças é Δw_{ij}, então a razão de chances correta seria $\exp(\Delta w_{ij}\cdot\gamma)$, e não $\Delta w_{ij}\cdot\exp(\gamma)$, pois esta última forma está incorreta.

b) Comparação entre idades distintas

A estimativa da razão de chances de preferência por Kcola correspondente à comparação de crianças do mesmo gênero x, com idades de 10 e 15 anos é dada por:

$$\log\left[\frac{\frac{\pi_i(x,10)}{1-\pi_i(x,10)}}{\frac{\pi_j(x,15)}{1-\pi_j(x,15)}}\right] = \log\left[\frac{\pi_i(x,10)}{1-\pi_i(x,10)}\right] - \log\left[\frac{\pi_j(x,15)}{1-\pi_j(x,15)}\right]$$

$$\log\left[\frac{\frac{\pi_i(x,10)}{1-\pi_i(x,10)}}{\frac{\pi_j(x,15)}{1-\pi_j(x,15)}}\right] = [\alpha + \beta x + \gamma(10-5)] - [\alpha + \beta x + \gamma(15-5)]$$

$$\log\left[\frac{\frac{\pi_i(x,10)}{1-\pi_i(x,10)}}{\frac{\pi_j(x,15)}{1-\pi_j(x,15)}}\right] = \alpha + \beta x + 5\gamma - \alpha - \beta x - 10\gamma$$

$$\log\left[\frac{\frac{\pi_i(x,10)}{1-\pi_i(x,10)}}{\frac{\pi_j(x,15)}{1-\pi_j(x,15)}}\right] = -5\gamma$$

$$\frac{\frac{\pi_i(x,10)}{1-\pi_i(x,10)}}{\frac{\pi_j(x,15)}{1-\pi_j(x,15)}} = \exp(-5\gamma) = \exp(-5 \times -0.03) = \exp(0.15) \approx 1.1618$$

c) Intervalos de Confiança

$\Longrightarrow$ Intervalo de Confiança de 95% para a razão de chances associada a β:

O intervalo é calculado como $\text{IC}_{95\%} = \exp(\hat{\beta} \pm z_{95\%} \times s_{\hat{\beta}})$, em que $s_{\hat{\beta}}$ é o erro padrão de $\hat{\beta}$.

$\text{IC}_{95\%} = \exp(0.33 \pm 1.96 \times 0.10) = [1.14; 1.69]$, que, como afirmado anteriormente, não inclui o valor um (que corresponderia a zero na escala logarítmica).

Com 95% de confiança, o intervalo $[1.14; 1.69]$ contém a verdadeira razão de chances de preferência por Kcola comparando uma criança do gênero masculino com uma do gênero feminino, mantendo as idades fixadas. Ou seja, se o estudo fosse repetido várias vezes utilizando o mesmo procedimento de amostragem, 95% dos intervalos de confiança construídos conteriam a verdadeira razão de chances populacional.

$\Longrightarrow$ Intervalo de Confiança de 95% para a razão de chances associada a γ:

Calculado como $\text{IC}_{95\%} = \exp(\hat{\gamma} \pm z_{95\%} \times s_{\hat{\gamma}})$, em que $s_{\hat{\gamma}}$ é o erro padrão.

$\text{IC}_{95\%} = \exp(-0.03 \pm 1.96 \times 0.005) = [0.96; 0.98]$.

Com 95% de confiança, o intervalo $[0.96; 0.98]$ contém a verdadeira razão de chances de preferir Kcola entre duas crianças do mesmo gênero, que diferem na idade em apenas um ano, sendo a primeira mais velha e a segunda mais nova. Da mesma forma, sob amostragens repetidas, 95% dos intervalos construídos conteriam esse verdadeiro parâmetro populacional.

d) Estimativa de Probabilidade

A probabilidade de uma criança do gênero x_i e idade w_i preferir Kcola pode ser estimada por $\pi_i(x_i, w_i)$, que é obtido algebricamente da seguinte forma:

$$
\begin{aligned}
\log\left[\frac{\pi_i(x_i,w_i)}{1-\pi_i(x_i,w_i)}\right] &= \alpha+\beta x_i+\gamma(w_i-5)\\
\frac{\pi_i(x_i,w_i)}{1-\pi_i(x_i,w_i)} &= \exp[\alpha+\beta x_i+\gamma(w_i-5)]\\
\pi_i(x_i,w_i) &= [1-\pi_i(x_i,w_i)]\exp[\alpha+\beta x_i+\gamma(w_i-5)]\\
\pi_i(x_i,w_i) &= \exp[\alpha+\beta x_i+\gamma(w_i-5)]-\pi_i(x_i,w_i)\exp[\alpha+\beta x_i+\gamma(w_i-5)]\\
\pi_i(x_i,w_i)+\pi_i(x_i,w_i)\exp[\alpha+\beta x_i+\gamma(w_i-5)] &= \exp[\alpha+\beta x_i+\gamma(w_i-5)]\\
\pi_i(x_i,w_i)[1+\exp(\alpha+\beta x_i+\gamma(w_i-5))] &= \exp[\alpha+\beta x_i+\gamma(w_i-5)]\\
\pi_i(x_i,w_i) &= \frac{\exp[\alpha+\beta x_i+\gamma(w_i-5)]}{1+\exp[\alpha+\beta x_i+\gamma(w_i-5)]} \quad (*)
\end{aligned}
$$

Assim, pela expressão $(*)$, a probabilidade de meninos ($x_i = 1$) com 15 anos preferirem Kcola é dada por:

$$
\pi_i(1,15) = \frac{\exp[0.69+0.33\times 1-0.03(15-5)]}{1+\exp[0.69+0.33\times 1-0.03(15-5)]} = \frac{\exp(0.72)}{1+\exp(0.72)} \approx 0.67 \text{ ou } 67\%
$$

Referências

1. Kvalseth, Tarald O. **Coefficient of variation: the second-order alternative.** Journal of Applied Statistics 44.3 (2017): 402-415.
2. Morettin, P. A. e Singer, J. M. **Estatística e Ciências de Dados**. Rio de Janeiro:LTC, 2023. (Já existe segunda edição do recente do livro, caso haja interressado pode comprar pelo link
3. Mazarei, A., Sousa, R., Mendes-Moreira, J., Molchanov, S., & Ferreira, H. M. (2025). **Online boxplot derived outlier detection.** International journal of data science and analytics, 19(1), 83-97.

MAE 5905: Introdução à Ciência de Dados

Lista 1. Primeiro Semestre de 2026. Entregar 06/04/2026.

1. Num conjunto de dados, o primeiro quartil é 10, a mediana é 15 e o terceiro quartil é 20. Indique quais das seguintes afirmativas são verdadeiras, justificando sua resposta:

0,25 v a) A distância interquartis é 5.

0,50 v b) O valor 32 seria considerado *outlier* segundo o critério utilizado na construção do *boxplot*.

0,50 v c) A mediana ficaria alterada de 2 unidades se um ponto com valor acima do terceiro quartil fosse substituído por outro 2 vezes maior.

0,25 v d) O valor mínimo é maior do que zero.

2. Num estudo na área de Oncologia, o número de vasos que alimentam o tumor está resumido na seguinte tabela.

Tabela 1: Distribuição de frequências do número de vasos que alimentam o tumor

Número de vasos	Frequência
$0 \vdash 5$	8 (12%)
$5 \vdash 10$	23 (35%)
$10 \vdash 15$	12 (18%)
$15 \vdash 20$	9 (14%)
$20 \vdash 25$	8 (12%)
$25 \vdash 30$	6 (9%)
Total	66 (100%)

1 v

Indique a resposta correta.

a) O primeiro quartil é 25%.

b) A mediana está entre 10 e 15.

c) O percentil de ordem 10% é 10.

d) A distância interquartis é 50.

e) Nenhuma das respostas anteriores.

3. Em um teste de esforço cardiopulmonar aplicado a 55 mulheres e 104 homens, foram medidas entre outras, as seguintes variáveis:

Tabela 2: VO2MAX

Grupo	n	Média	Mediana	Desvio Padrão
Normais	56	1845	1707	795
Cardiopatas	57	1065	984	434
DPOC	46	889	820	381

Tabela 3: VCO2MAX

Grupo	n	Média	Mediana	Desvio Padrão
Normais	56	2020	1847	918
Cardiopatas	57	1206	1081	479
DPOC	46	934	860	430

- Grupo: Normais, Cardiopatas ou DPOC (portadores de doença pulmonar obstrutiva crônica).
- VO2MAX: consumo máximo de O2 (ml/min).
- VCO2MAX: consumo máximo de CO2 (ml/min).

Algumas medidas descritivas e gráficos são apresentados abaixo nas Tabelas 2 e 3 e Figura 1 Coeficiente de correlação entre VO2MAX e VCO2MAX = 0,92.

1,5 v a) Que grupo tem a maior variabilidade?

0,25v b) Compare as médias e as medianas dos 3 grupos.

0,50 v c) Compare as distâncias interquartis dos 3 grupos para cada variável. Você acha razoável usar a distribuição normal para esse conjunto de dados?

0,25 v d) O que representam os asteriscos nos *boxplots*?

0,25 v e) Que tipo de função você ajustaria para modelar a relação entre o consumo máximo de CO2 e o consumo máximo de O2? Por quê?

0,25 v f) Há informações que necessitam verificação quanto a possíveis erros? Quais?

1,0 v **4.** O gráfico QQ da Figura 2 corresponde ao ajuste de um modelo de regressão linear múltipla.

Pode-se afirmar que:

a) Há indicações de que a distribuição dos erros é Normal.

b) Há evidências de que a distribuição dos erros é assimétrica.

c) Há evidências de que a distribuição dos erros tem caudas mais leves do que aquelas da distribuição Normal.

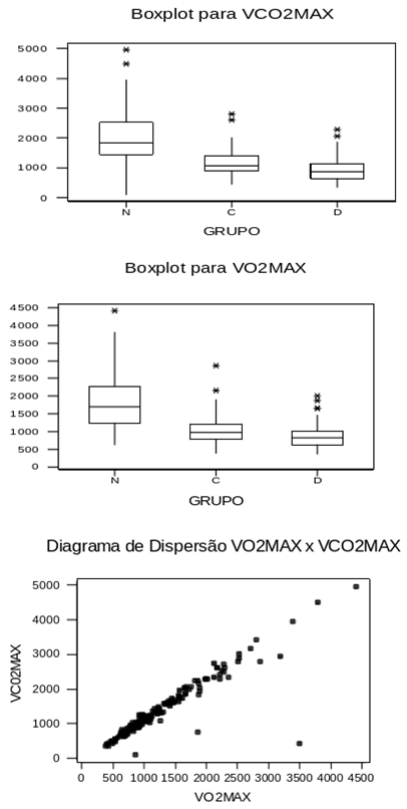


Figura 1: Gráficos para o Exercício 3.

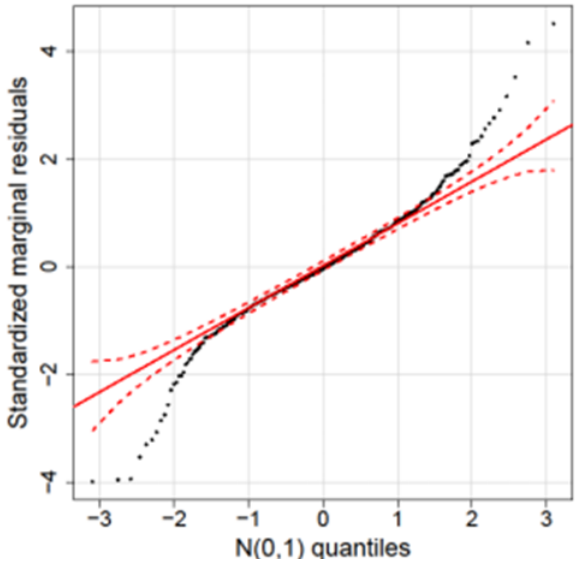


Figura 2: Gráfico QQ correspondente ajuste de um modelo de regressão linear múltipla.

d) Há evidências de que a distribuição dos erros tem caudas mais pesadas que aquelas da distribuição Normal.

e) Nenhuma das anteriores.

5. Para estudar a associação entre gênero (1=Masc, 0=Fem) e idade (anos) e a preferência (1=sim, 0=não) pelo refrigerante Kcola, o seguinte modelo de regressão logística foi ajustado aos dados de 50 crianças escolhidas ao acaso:

$$\log\left\{\frac{\pi_i(x_i, w_i)}{1-\pi_i(x_i, w_i)}\right\} = \alpha + \beta x_i + \gamma(w_i - 5),$$

em que x_i (w_i) representa o gênero (idade) da i-ésima criança e $\pi_i(x_i,w_i)$ a probabilidade de uma criança do gênero x_i e idade w_i preferir Kcola. As seguintes estimativas para os parâmetros foram obtidas:

Parâmetro	Estimativa	Erro padrão	Valor p
α	0,69	0,12	$< 0{,}01$
β	0,33	0,10	$< 0{,}01$
γ	-0,03	0,005	$< 0{,}01$

1,5 v a) Interprete os parâmetros do modelo por intermédio de chances e razões de chances.

0,50 v b) Com as informações acima, estime a razão de chances de preferência por Kcola correspondente à comparação de crianças do mesmo gênero com 10 e 15 anos.

1 v c) Construa intervalos de confiança (com coeficiente de confiança aproximado de 95%) para $\exp(\beta)$ e $\exp(\gamma)$ e traduza o resultado em linguagem não técnica.

0,50 d) Estime a probabilidade de meninos com 15 anos preferirem Kcola.