

Quinta Lista de Exercícios: Análise de Regressão e Agrupamento Hierárquico

Exercício 1. Na estação de pesquisa Palmer na **Antártica 1**, pesquisadores fizeram medições em três espécies diferentes de pinguins: Adélie, Chinstrap e Gentoo. Os dados obtidos estão no arquivo **penguins_size.csv** (pasta Dados na seção Arquivos no Microsoft Teams). Salve o arquivo em seu diretório de trabalho e, em seguida, guarde-o dentro da variável **penguins**:

```
#lendo um arquivo do Excel que possui cabeçalho
penguins <- read.csv("penguins_size.csv", header = TRUE)
```

Vamos inicialmente conhecer o conjunto de dados. Para isso, comece utilizando o comando **str(penguins)**. A partir da saída desse comando, já é possível afirmar que o conjunto de dados possui 344 observações e cada observação possui 7 variáveis, como exemplo, sexo, espécie, ilha onde o pinguim habita, peso em gramas (**body_mass_g**), tamanho da asa em milímetros (**flipper_length_mm**), dentre outras. Agora que já compreendemos um pouco melhor o nosso conjunto de dados, podemos dar início a nossa análise estatística. Os primeiros exercícios referem-se à análise descritiva dos dados. Os últimos exercícios referem-se à análise de regressão.

- Determine o peso médio e o comprimento médio da asa dos pinguins. Se você encontrou algum erro ao utilizar a função **mean()**, talvez seja porque há dados faltantes! utilize o argumento **na.rm = TRUE** para corrigir esse erro (esse argumento exclui os dados faltantes para, em seguida, realizar os cálculos da função). Agora, calcule o desvio padrão de cada uma dessas variáveis (função **sd()**).
- Refaça a parte (a), mas agora calculando a média e o desvio padrão do peso e do comprimento da asa para cada uma das espécies. Comente os resultados encontrados.
- Refaça a parte (a), mas agora calculando a média e o desvio padrão do peso e do comprimento da asa para cada um dos sexos. Comente os resultados encontrados.
- Plote o gráfico de **flipper_length_mm** versus **body_mass_g**.
- Utilize a função **cor()** para calcular a correlação entre as variáveis **flipper_length_mm** e **body_mass_g**.
- A partir das duas respostas anteriores, responda: há alguma relação entre **flipper_length_mm** e **body_mass_g**? Quão forte é essa relação? A relação é positiva ou negativa?
- Utilize a função **lm()** para determinar a reta do modelo de regressão linear simples. Considere **flipper_length_mm** como a variável explanatória (x) e **body_mass_g** como a variável resposta (y).
- Explique o coeficiente angular da reta encontrada em (g).

- (i) A partir do modelo linear encontrado em (g), qual seria o peso médio de um pinguim que possui uma asa de 204 mm? Você poderia utilizar esse modelo para estimar o peso médio de um pinguim que tivesse uma asa de 168 mm? Justifique sua resposta.

Exercício 2. Considere a seguinte matriz de distâncias:

$$\begin{bmatrix} 0 & 9 & 3 & 6 & 11 \\ 9 & 0 & 7 & 5 & 10 \\ 3 & 7 & 0 & 9 & 2 \\ 6 & 5 & 9 & 0 & 8 \\ 11 & 10 & 2 & 8 & 0 \end{bmatrix}$$

- (a) Com base na matriz de distâncias acima, esboce o dendograma que resulta do processo de aglomeração hierárquica dessas 5 observações usando o método **complete** como a distância entre dois aglomerados.
- (b) Repita o exercício (a) utilizando o método **single** como a distância entre dois aglomerados.
- (c) Suponha que um corte seja feito no dendograma encontrado em (a) de forma a deixar dois aglomerados. Quais observações estão em cada aglomerado?
- (d) Suponha que um corte seja feito no dendograma encontrado em (b) de forma a deixar dois aglomerados. Quais observações estão em cada aglomerado?