Trabalho AEII – 2023/24

1. O conjunto de dados “penguins” do package “palmerpenguins” inclui medidas para três espécies de pinguins (Adélie, Chinstrap e Gentoo) da ilha no Arquipélago Palmer, relativas a comprimento das barbatanas, massa corporal, dimensões do bico e sexo. O conjunto de dados contém 8 variáveis para 344 pinguins. Neste trabalho vamos utilizar apenas 3 variáveis: espécie (species), sexo (sex) e massa corporal do pinguim em gramas (body_mass_g).

body_mass_g é a variável quantitativa contínua e será a variável dependente, enquanto espécie e sexo são variáveis qualitativas. Certifique-se de que as variáveis qualitativas são consideradas como fatores para o R. Caso contrário, precisarão ser transformadas em fatores.

Antes de executar qualquer teste, faça uma análise de resíduos. Seguidamente analise primeiro os dados utilizando gráficos de interação, que ajudarão a visualizar o efeito de um fator, através dos níveis de outro fator. Execute o modelo de análise de variância que lhe parecer mais conveniente utilizando todos os recursos necessários para a análise dos dados. Considere um nível de significância de 5%.

Recordar que:

- O teste F da ANOVA e as comparações múltiplas de Tukey são relativamente robustos a desvios à hipótese de normalidade.
- As violações ao pressuposto de variâncias homogéneas são em geral menos graves no caso de delineamentos equilibrados, mas podem ser graves em delineamentos não equilibrados.

Para testar a normalidade dos dados entre diferentes células podem usar o comando
aggregate(body_mass_g ~ species *sex, data= penguins, function(x) shapiro.test(x)$p.value)

A hipótese de homogeneidade de variâncias entre diferentes células pode ser testada recorrendo ao Teste de Bartlett, caso a dimensão da amostra seja grande (e.g., $n_{ij} \geq 5$ em todas as células).

Assume-se que os erros são independentes.

Para podem também usar a função aggregate para calcular as médias por espécie e por sexo:
aggregate(body_mass_g ~ species *sex, data= penguins, FUN=mean)

2. O conjunto de dados “sat” do package “faraway” foi obtido com o objetivo de estudar a relação entre as despesas dos alunos com a educação no ensino público e os resultados obtidos no exame SAT. O conjunto de dados contém 7 variáveis relativas aos resultados de 50 alunos. Considere como variável resposta a pontuação média total no SAT (total) e considere como preditores as variáveis: despesas (expend), ordenado (salary), a razão média de alunos por professor (ratio) e a percentagem de alunos elegíveis para fazerem o exame (takers).

Ajuste um modelo de regressão linear múltipla aos dados e teste a sua significância, $\alpha = 0,05$. Comece por fazer uma análise de resíduos ao modelo completo. Realize todos os testes que achar relevantes para o conjunto de dados em análise. Utilize a função stepwise para obter o modelo que “melhor” se ajusta ao conjunto de dados “sat”. Utilize ainda testes F-parciais para confirmar o resultado obtido pela função stepwise. Explore da maneira que achar conveniente as funções do R para a regressão linear múltipla.

Nota: A entrega do trabalho deve ser feita até ao final do dia 20 de dezembro de 2023. Devem entregar um relatório com a resolução das questões propostas. As apresentações do trabalho serão no dia 22 de dezembro de 2023.