INSTITUTO SUPERIOR DE AGRONOMIA

ESTATÍSTICA E DELINEAMENTO - 2023-24

8 janeiro 2024 **Primeira Chamada de Exame** **SOLUÇÕES**

I

1. O melhor modelo de regressão linear simples para modelar a variável `WRV` é com a variável `Thick` pois é a que tem o maior valor do coeficiente de correlação, em módulo, com `WRV` ($|r_{x,y}| = 0.7340$).

2. (a) $b_1 = -133.1932$. Indica que a retenção de água (variável `WRV`) diminui em média 133.1932% quando a espessura da folha (variável `Thick`) aumenta 1 mm.

   (b) No teste de ajustamento global, rejeita-se H_0: $\mathcal{R}^2 = 0$ a favor de H_1: $\mathcal{R}^2 > 0$, ao nível de significância $\alpha = 0.05$, ou seja, o modelo ajustado é não inútil. No entanto, tendo em conta o valor do $R^2 = 0.539$, este modelo só explica cerca de 54% da variabiliade total observada da variável `WRV`.

   (c) Demonstração feita nas aulas teóricas da UC e apresentada na página 40 dos Apontamentos de Estatística e Delineamento, O Modelo Linear (Cadima, 2020).

   (d) $\hat{\sigma}_{\hat{\beta}_1} = 38.9719$

   (e) $]-220.028, -46.3584[$
   Este intervalo representa, com 95% de confiança, a gama de valores admissíveis para β_1, o valor populacional do declive da recta de regressão.

3. Resolução feita nas aulas teóricas da UC e apresentada na página 19 dos Apontamentos de Estatística e Delineamento, O Modelo Linear (Cadima, 2020).

II

1. X, a matriz do modelo, tem $n = 12$ linhas e $p + 1 = 5$ colunas,

$$X = \begin{bmatrix} \vec{1}_n & \overrightarrow{\texttt{Gram}} & \overrightarrow{\texttt{Thick}} & \overrightarrow{\texttt{Smooth}} & \overrightarrow{\texttt{Por}} \\ | & | & | & | & | \end{bmatrix},$$

   em que, $\vec{1}_n$ é um vector com 12 uns e as restantes colunas são os vectores com os 12 valores observados de cada uma das quatro variáveis preditoras.

2. Simbolicamente, a pergunta é $\beta_3 = \beta_4$? Com base no teste t a uma combinação linear de βs, de hipóteses $H_0 : \beta_3 - \beta_4 = 0 \quad vs. \quad H_1 : \beta_3 - \beta_4 \neq 0$, com um nível de significância $\alpha = 0.05$, não se rejeita a hipótese nula, sendo admissível considerar que o valor populacional do coeficiente do preditor `Smoth` é igual ao valor populacional do coeficiente do preditor `Por` (nota: $T_{cal} = \frac{-0.6072-(-0.3387)-0}{\sqrt{5.2101+0.1383-2\times0.3734}} = -0.125167$).

3. (a) Dedução incluída nas demonstrações que foram feitas nas aulas teóricas e práticas.

   (b) $R^2_{mod} = 0.410557$. Como era suposto, $R^2_{mod} < R^2$. No entanto, o facto de R^2_{mod} ser muito inferior a R^2 fica a dever-se ao número de observações ($n = 12$) não ser muito maior que o número de parâmetros do modelo ($p + 1 = 5$), resultando numa forte penalização do já não muito grande R^2. Explicação feita nas aulas teóricas e apresentada nas páginas 145 e 146 dos Apontamentos de Estatística e Delineamento, O Modelo Linear (Cadima, 2020).

4. (a) A primeira variável a ser excluída é `Smooth` pois, considerando as variáveis que apresentam um *p-value* superior a $\alpha = 0.1$ para o teste t de hipóteses $H_0 : \beta_i = 0 \quad vs. \quad H_1 : \beta_i \neq 0$ $(i = 1, \ldots, 4)$, é a que tem maior valor de prova neste teste, *p-value*= 0.7979.

   (b) Usando o algoritmo de exclusão sequencial com base no AIC, o submodelo final tem apenas o preditor `Thick` (o modelo com menor AIC). De acordo com o teste F parcial, o submodelo final com um só preditor (`WRV` $\sim$ `Thick`), não difere significativamente do modelo original com 4 preditores, ao nível de significância $\alpha = 0.05$, pelo que se deve preferir o submodelo por ser mais simples (nota: $F_{Parcial_{calc}} = 0.5358642$).

5. (a) A explicação sobre a interpretação destes gráficos está nas páginas 72 a 75 dos Apontamentos de Estatística e Delineamento, O Modelo Linear (Cadima, 2020), assim como nos slides das aulas teóricas.

   - Gráfico de resíduos (usuais), eixo vertical, contra valores ajustados $\hat{y}_i$, no eixo horizontal: observa-se uma ligeira curvatura sugerindo que o modelo de regressão linear simples pode não ser a melhor forma de relacionar o rendimento da pasta com a lenhina residual na pasta. Não se visualiza uma forma de funil sendo de admitir válido o pressuposto da homogeneidade das variâncias dos erros aleatórios. O resíduo correspondente à observação 5 surge destacado dos restantes, os quais se dispõem numa banda oblíqua, indicando a existência de possíveis observações atípicas.
   - qq-plot, gráfico dos quantis empíricos dos resíduos estandardizados, eixo vertical, vs quantis teóricos da $N(0, 1)$, eixo horizontal: embora surja um certo afastamento da linearidade nos três pontos mais à direita, dado o reduzido número de observações da amostra, a hipótese de normalidade dos erros aleatórios parece ser admissível.
   - Gráfico de resíduos internamente estandardizados, eixo vertical, vs valores do efeito alavanca, eixo horizontal: destaca-se a observação 5 com o maior valor absoluto de resíduo estandardizado e o maior valor do efeito alavanca, bastante superior ao valor médio de $\frac{2}{12}$, sugerindo que esta observação está a atrair a reta ajustada. O valor da distância de Cook da observação 5 é superior a 1 ($>> 0.5$) indicando que estamos na presença de uma observação muito influente (isto é, uma observação que, se retirada do conjunto de dados, gera uma mudança assinalável nos resultados do ajustamento).

   (b) No teste t, de hipóteses $H_0 : \beta_2 = 0 \quad vs. \quad H_1 : \beta_2 \neq 0$, rejeita-se a hipótese nula para qualquer um dos usuais níveis de significância (*p-value*= $0.000704 << \alpha$). Logo, o modelo parabólico ajusta-se significativamente melhor que o correspondente modelo de regressão linear simples.

III

1. (a) Delineamento fatorial a 2 fatores, clone (7 níveis) e quinta (5 níveis). Existem 8 repetições em cada uma das 35 situações experimentais pelo que se trata de um delineamento equilibrado com um total de 280 observações. Havendo repetições para cada situação experimental, a interação entre factores deverá ser estudada. O modelo correspondente a este delineamento, será o modelo ANOVA a 2 fatores de efeitos fixos com interação, equilibrado, descrito na página 198 das folhas de apoio, assim como nos slides das aulas teóricas. Trata-se de adaptar este modelo ao problema em estudo.

   (b) Coluna `Df`, linha `Clone`: 6;
   Coluna `Df`, linha `Quinta`: 4;
   Coluna `Df`, linha `Clone:Quinta`: 24;

Coluna `Df`, linha `Residuals`: 245;
Coluna `Mean Sq`, linha `Clone`: 15.683;
Coluna `F Value`, linha `Clone:Quinta`: 0.3276.

(c) Fazendo os testes F aos efeitos de interação, aos efeitos principais do factor Clone e aos efeitos principais do factor Quinta, conclui-se, ao nível de significância $\alpha = 0.05$, que não existem efeitos de interação entre níveis dos 2 factores e que existem efeitos principais do factor Clone e do factor Quinta.

(d) Responde-se à questão efetuando um Teste de Tukey. Conclui-se que a afirmação é verdadeira, isto é, ao nível de significância $\alpha = 0.05$, os rendimentos médios observados do clone `AV1` nas Quintas `QT4` e `QT5` não são significativamente diferentes (Termo de comparação de Tukey (TC), $TC = q_{0.05(35,245)} \sqrt{\frac{5.8}{8}}$).

(e) Resultado demonstrado nas aulas teóricas e na página 208 dos Apontamentos de Estatística e Delineamento, O Modelo Linear (Cadima, 2020).

2. De acordo com o delineamento experimental descrito (delineamento experimental em blocos completos casualizados), o teste não paramétrico a realizar é o teste de Friedman. Tem-se 2 fatores, clone ($a = 7$ níveis) e bloco ($b = 6$ níveis), em que o segundo factor (bloco) pode afectar a variável resposta (número de pegamentos em 100 enxertias) e se deseja controlar os seus efeitos (efeitos dos diferentes enxertadores).

   H_0: Os valores observados relativos ao pegamento da enxertia em cada clone são globalmente semelhantes, tendo em conta eventuais diferenças em diferentes blocos (enxertadores),
   vs.
   H_1: Há clones com valores observados relativos ao pegamento da enxertia globalmente maiores (ou menores) do que outros, tendo em conta eventuais diferenças em diferentes blocos (enxertadores).

   $S_{calc} = 22.4$, $\chi^2_{0.05(6)} = 12.592$, logo, rejeita-se H_0 para $\alpha = 0.05$. Para este nível de significância, há evidência estatística para considerar que há clones com pegamentos da enxertia globalmente maiores (ou menores) do que outros, tendo em conta eventuais diferenças em diferentes blocos (enxertadores).