

INSTITUTO SUPERIOR DE AGRONOMIA
ESTATÍSTICA E DELINEAMENTO – 2023-24

8 janeiro 2024 Primeira Chamada de Exame Duração: 3h00 (+15min)

I [6 valores]

Num estudo realizado com várias espécies de eucalipto, efectuaram-se 12 medições independentes, em árvores selecionadas ao acaso com 6 anos de idade, das seguintes características relativas às propriedades da folha do papel: retenção de água (variável **WRV**, em %) gramagem da folha (variável **Gram**, em g/m^2); espessura da folha (variável **Thick**, em mm); lisura da folha (variável **Smooth**, em l/min); permeabilidade ao ar da folha (variável **Por**, em l/min). Eis os seguintes indicadores, bem como a matriz de correlações para todos os pares de variáveis.

Variável	Mínimo	Média	Máximo	Variância
WRV	49.82	54.53	57.82	6.5857
Gram	59.72	63.31	64.77	1.7281
Thick	0.0860	0.1098	0.1360	0.0002
Smooth	0.1860	0.5922	1.1710	0.1013
Por	0.3710	2.0251	6.6930	3.4829

```
> round(cor(papel),4)
           WRV      Gram      Thick      Smooth      Por
WRV      1.0000 -0.0352 -0.7340 -0.0624 -0.3525
Gram     -0.0352  1.0000  0.3089 -0.3880  0.2300
Thick    -0.7340  0.3089  1.0000  0.0149  0.2579
Smooth   -0.0624 -0.3880  0.0149  1.0000 -0.4433
Por      -0.3525  0.2300  0.2579 -0.4433  1.0000
```

1. Qual o melhor modelo de regressão linear simples para modelar a retenção de água (variável **WRV**)? Justifique a sua resposta.
2. Para a regressão linear simples escolhida na alínea anterior:
 - (a) Determine o declive da recta ajustada e interprete o seu valor.
 - (b) Descreva, em pormenor, o teste F de ajustamento global. Para a regressão ajustada o que conclui?
 - (c) Dado o Modelo de Regressão Linear Simples e sabendo que $\hat{\beta}_1 = \frac{cov_{xY}}{s_x^2} = \sum_{i=1}^n c_i Y_i$, com $c_i = \frac{x_i - \bar{x}}{(n-1)s_x^2}$, prove que a variância do estimador do declive da recta populacional é

$$V[\hat{\beta}_1] = \frac{\sigma^2}{(n-1)s_x^2}.$$

- (d) Calcule a estimativa do desvio padrão associada ao estimador do declive.
- (e) Indique o intervalo a 95% confiança para o parâmetro β_1 . Interprete o resultado obtido.
3. Uma curva *exponencial* é uma curva com a seguinte equação:

$$y = c e^{dx}$$

com $y > 0$; $c > 0$. Linearize este modelo explicitando os seus parâmetros.

II [7 valores]

Para o mesmo conjunto de dados descritos no grupo I, foi seguidamente ajustado um modelo de regressão linear múltipla para modelar a retenção de água (variável *WRV*) a partir das restantes variáveis. Eis os resultados obtidos com o ajustamento deste modelo, bem como a matriz de (co-) variâncias estimadas dos estimadores dos parâmetros do modelo:

Call:

```
lm(formula = WRV ~ ., data = papel)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	42.1748	32.2890	1.306	0.2328
Gram	0.4296	0.5220	0.823	0.4376
Thick	-125.7063	43.8702	-2.865	0.0241
Smooth	-0.6072	2.2826	-0.266	0.7979
Por	-0.3387	0.3719	-0.911	0.3926

Residual standard error: 1.97 on 7 degrees of freedom

Multiple R-squared: 0.6249, Adjusted R-squared: ?

F-statistic: 2.916 on 4 and 7 DF, p-value: 0.103

```
> round(vcov(papel.lm), 4)
```

	(Intercept)	Gram	Thick	Smooth	Por
(Intercept)	1042.5769	-16.6905	301.6822	-30.0720	-0.4017
Gram	-16.6905	0.2725	-7.7072	0.4593	0.0066
Thick	301.6822	-7.7072	1924.5977	-25.9721	-4.7181
Smooth	-30.0720	0.4593	-25.9721	5.2101	0.3734
Por	-0.4017	0.0066	-4.7181	0.3734	0.1383

1. Para o modelo ajustado, descreva a matriz do modelo (\mathbf{X}).
2. Será admissível considerar que o correspondente valor populacional do coeficiente do preditor *Smooth* é igual ao valor populacional do coeficiente do preditor *Por*? Efectue um teste de hipóteses adequado para avaliar esta hipótese ($\alpha=0.05$).
3. O R^2 modificado, $R_{mod}^2 = 1 - \frac{QMRE}{QMT}$, é um indicador importante a ter em conta numa análise de regressão linear múltipla:

(a) Prove que

$$R_{mod}^2 = 1 - (1 - R^2) \cdot \frac{n-1}{n-(p+1)},$$

sendo R^2 o coeficiente de determinação usual.

(b) Para o modelo de regressão linear múltipla ajustado, calcule o R^2 modificado e comente o valor obtido.

4. O técnico responsável pelo estudo, não satisfeito com o resultado obtido, decidiu aplicar um algoritmo de seleção de variáveis.
 - (a) Num algoritmo de exclusão sequencial com base nos testes t aos coeficientes dos preditores, com nível de significância $\alpha=0.10$, qual a primeira variável a ser excluída? Justifique a sua resposta.
 - (b) O algoritmo de exclusão sequencial com base no Critério de Informação de Akaike (*AIC*) produziu os seguintes resultados finais:

```

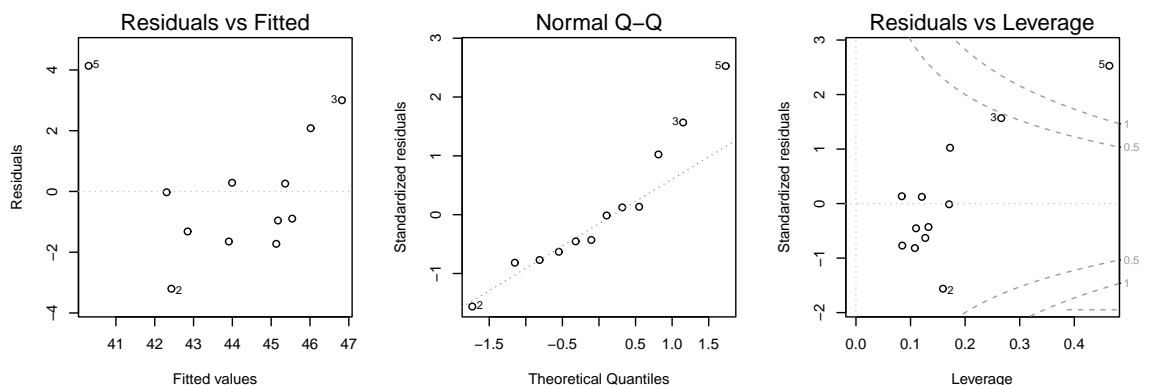
> step(papel.lm)
...
Step: AIC=17.19
WRV ~ Gram + Thick
      Df Sum of Sq  RSS   AIC
- Gram  1     2.940 33.421 16.291
<none>                30.482 17.187
- Thick 1    41.878 72.359 25.561
Step: AIC=16.29
WRV ~ Thick
      Df Sum of Sq  RSS   AIC
<none>                33.421 16.291
- Thick 1    39.028 72.449 23.576

```

Teste formalmente, com nível de significância $\alpha = 0.05$, se o ajustamento do submodelo final difere significativamente de modelo original com 4 preditores.

5. No estudo em curso com as várias espécies de eucalipto, efectuaram-se ainda 12 medições independentes de duas características relativas à qualidade da pasta de papel: rendimento da pasta e lenhina residual na pasta.

- (a) Primeiramente foi sugerido ajustar uma regressão linear simples do rendimento da pasta sobre a lenhina residual na pasta. Os gráficos dos resíduos obtidos com o ajustamento desse modelo de regressão linear simples apresentam-se seguidamente:



Analise os 3 gráficos e discuta as suas implicações para o estudo efetuado.

- (b) Seguidamente foi sugerido modelar o rendimento da pasta a partir do ajustamento de um polinómio de segundo grau. Eis os resultados obtidos:

```

Call:
lm(formula = rendimento ~ lenhina + I(lenhina^2), data = pasta)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  93.48021    8.24318   11.340 1.24e-06
lenhina      -5.29534    0.95348   -5.554 0.000355
I(lenhina^2)  0.13463    0.02673    5.036 0.000704
---
Residual standard error: 1.206 on 9 degrees of freedom
Multiple R-squared:  0.8529, Adjusted R-squared:  0.8203
F-statistic: 26.1 on 2 and 9 DF, p-value: 0.0001794

```

Com base na informação disponível, responda sucintamente se o modelo parabólico se ajusta de forma significativamente melhor que o modelo de regressão linear simples de rendimento da pasta sobre lenhina residual na pasta ($\alpha=0.05$).

III [7 valores]

- Num estudo sobre clones selecionados da variedade de videira Avesso conduzido em Cinfães, foram estudados 7 clones, designados AV1, AV2, AV3, AV4, AV5, AV6 e AV7, em 5 diferentes Quintas, designadas QT1, QT2, QT3, QT4 e QT5. No ensaio realizado em cada Quinta, foram aleatoriamente associadas 8 parcelas a cada clone e avaliou-se o rendimento das plantas por parcela (kg/planta). As médias amostrais por combinação Clone:Quinta são indicadas seguidamente:

Clone:Quinta		Quinta				
Clone	QT1	QT2	QT3	QT4	QT5	
AV1	1.787	4.254	5.546	10.715	9.432	
AV2	1.107	3.485	5.079	7.908	8.743	
AV3	0.858	3.551	4.678	8.598	7.900	
AV4	1.463	3.869	7.042	9.379	8.756	
AV5	1.256	2.936	6.055	8.984	8.449	
AV6	0.207	1.706	4.562	8.326	7.668	
AV7	1.130	3.595	5.514	9.353	9.111	

- Descreva em pormenor o delineamento experimental e o respetivo modelo ANOVA adequado ao problema sob estudo.
- Seguidamente apresenta-se a tabela ANOVA relativa ao estudo efectuado:

```
> Avesso.aov<-aov(Rend~Clone*Quinta, data=Avesso)
> summary(Avesso.aov)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Clone	?	94.1	?	2.688	0.0152
Quinta	?	2569.1	642.3	110.101	<2e-16
Clone:Quinta	?	44.9	1.9	?	0.9992
Residuals	?	1429.2	5.8		

Complete a tabela, indicando como obtém cada um dos valores em falta.

- Quais os tipos de efeitos que devem ser testados? Caso necessite de vários testes, descreva um em pormenor e enuncie apenas as conclusões para os restantes (ao nível $\alpha=0.05$).
- Comente a seguinte afirmação: "ao nível de significância $\alpha=0.05$, os rendimentos médios observados do clone AV1 nas Quintas QT4 e QT5 não são significativamente diferentes".
- Prove que no contexto da ANOVA descrita neste exercício, o Quadrado Médio Residual é a média (simples) das variâncias amostrais de célula (S_{ij}^2) das situações experimentais (ab):

$$QMRE = \frac{1}{ab} \sum_{i=1}^a \sum_{j=1}^b S_{ij}^2$$

- No mesmo estudo com os 7 clones da variedade Avesso, os mesmos clones foram estudados relativamente ao número de pegamentos em 100 enxertias, num ensaio com um delineamento experimental em blocos completos casualizados, com 6 blocos (correspondentes a 6 diferentes

enxertadores). O responsável pelo estudo optou por uma análise de variância de tipo não paramétrico. Os dados da experiência, as respetivas ordenações do clone por bloco (entre parênteses), bem como a soma das ordenações do clone por bloco (Soma(ord): R_i , para $i = 1, \dots, 7$), encontram-se seguidamente descritos.

Bloco	Número de pegamentos (Ordenação)						
	Clone						
	AV7	AV2	AV3	AV4	AV5	AV6	AV1
Bloco1	35 (1)	58 (2)	71 (4)	88 (7)	80 (6)	74 (5)	60 (3)
Bloco2	17 (1)	44 (3)	41 (2)	61 (5)	70 (6)	56 (4)	80 (7)
Bloco3	35 (1)	77 (6)	63 (2)	67 (3)	70 (4)	75 (5)	80 (7)
Bloco4	47 (1)	65 (2)	66 (3)	64 (4)	75 (7)	67 (5)	69 (6)
Bloco5	35 (1)	70 (4)	47 (2)	55 (3)	75 (5)	77 (6)	94 (7)
Bloco6	44 (1)	45 (2)	80 (6)	79 (5)	50 (3)	69 (4)	81 (7)
Soma(Ord)	(6)	(19)	(19)	(27)	(31)	(29)	(37)

De acordo com o delineamento experimental do ensaio, o que conclui sobre o estudo efectuado ao nível $\alpha=0.05$? Descreva em pormenor o teste realizado.

Notas:

(a) Estatística do teste de Kruskal Wallis: sob H_0 ,

$$H = \frac{12}{n(n+1)} \sum_{i=1}^k \frac{R_i^2}{n_i} - 3(n+1) \sim \chi_{k-1}^2 ;$$

(b) Estatística do teste de Friedman: sob H_0 ,

$$S = \frac{12}{ab(a+1)} \sum_{i=1}^a R_i^2 - 3b(a+1) \sim \chi_{a-1}^2 ;$$

(c) $\chi_{0.05(5)}^2 = 11.070$; $\chi_{0.05(6)}^2 = 12.592$; $\chi_{0.05(7)}^2 = 14.067$.