INSTITUTO SUPERIOR DE AGRONOMIA

ESTATÍSTICA E DELINEAMENTO – 2020-21

12 Julho 2021 **Segunda Chamada de EXAME** **Duração: 3h00**

Aviso: Justifique convenientemente as suas respostas.

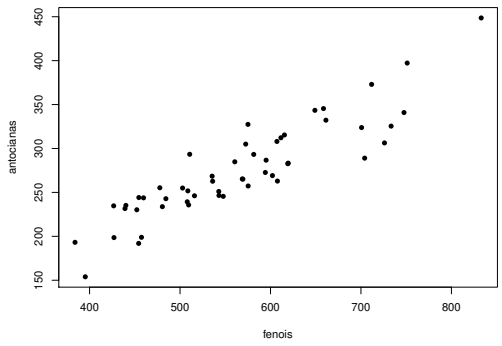
I [9 valores]

Um estudo pretende modelar o teor de antocianas (em mg/l) de bagos de uva da casta Moreto, uma casta tinta cultivada no Alentejo. Em $n=52$ genótipos, foram observados valores médios da variável `antocianas` e de cinco potenciais variáveis preditoras: teor de fenóis (em mg/l); teor brix (graus brix); pH; peso dos bagos (em g); e acidez (em g/l de ácido tartárico). São conhecidos os seguintes indicadores relativos às observações de cada variável, bem com a respectiva matriz de correlações:

	`antocianas`	`pH`	`brix`	`fenois`	`acidez`	`pesobago`
Mínimo	153.923	3.827	17.400	383.934	3.200	2.200
Média	275.0065	3.9703	19.6769	564.3342	3.7683	2.4697
Máximo	448.615	4.103	21.333	832.909	4.650	2.951
Desvio Padrão	53.6160	0.0593	0.6468	101.9808	0.2830	0.1357

	antocianas	pH	brix	fenois	acidez	pesobago
antocianas	1.00000	0.27786	0.65778	0.89735	-0.14286	-0.13566
pH	0.27786	1.00000	0.59785	0.32601	-0.49489	0.07244
brix	0.65778	0.59785	1.00000	0.60313	-0.22063	0.07903
fenois	0.89735	0.32601	0.60313	1.00000	-0.13315	-0.11272
acidez	-0.14286	-0.49489	-0.22063	-0.13315	1.00000	0.23733
pesobago	-0.13566	0.07244	0.07903	-0.11272	0.23733	1.00000

1. Em baixo vê-se o gráfico relacionando as variáveis `antocianas` e `fenois`.



   (a) Calcule a recta de regressão de `antocianas` sobre `fenois`.

   (b) Discuta em pormenor a qualidade do ajustamento da recta de regressão.

   (c) Será admissível afirmar que a cada mg por litro adicional no teor de fenóis corresponde, na população, um aumento médio de 0.5 mg/l no teor de antocianas? Responda através dum intervalo a 95% de confiança, sabendo que a estimativa da variância dos erros aleatórios do modelo é 571.186.

   (d) Calcule o resíduo usual associado à observação que se encontra no canto superior direito.

(e) Qual é a observação com o maior efeito alavanca? Calcule o respectivo valor.

2. Um analista considerou que a nuvem de pontos acima apresenta uma curvatura que justificaria o ajustamento dum polinómio de terceiro grau. Eis os resultados obtidos:

```
> summary(lm(antocianas ~ fenois + I(fenois^2) + I(fenois^3) , data=moretoEx))
[...]
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.114e+02  4.214e+02  -1.925   0.0601
fenois       4.906e+00  2.196e+00   2.234   0.0302
I(fenois^2) -7.789e-03  3.738e-03  -2.084   0.0425
I(fenois^3)  4.445e-06  2.080e-06   2.138   0.0377
---
Residual standard error: 23.2 on 48 degrees of freedom
Multiple R-squared:  0.8238,  Adjusted R-squared:  0.8128
F-statistic: 74.81 on 3 and 48 DF,  p-value: < 2.2e-16
```

(a) Escreva a equação da curva ajustada.

(b) Teste formalmente a hipótese deste modelo cúbico ter um ajustamento significativamente melhor que o modelo de regressão linear simples inicial. Comente.

3. Foi ajustada uma regressão linear múltipla com a totalidade dos preditores disponíveis, tendo-se obtido os seguintes resultados.

```
Call: lm(formula = antocianas ~ . , data = moretoEx)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  245.50139  276.42909   0.888  0.37910
pH          -147.20363   75.49169  -1.950  0.05729
brix          23.78472    7.24378   3.283  0.00196
fenois         0.40251    0.03928  10.247 1.86e-13
acidez        -8.80027   13.46178  -0.654  0.51655
pesobago     -19.45681   24.83780  -0.783  0.43743
---
Residual standard error: 22.31 on 46 degrees of freedom
Multiple R-squared: 0.8439,  Adjusted R-squared: 0.8269
F-statistic: 49.72 on 5 and 46 DF,  p-value: < 2.2e-16
```

(a) É possível afirmar que, na população, mantendo constantes os restantes preditores, a um aumento de acidez corresponde uma diminuição no teor de antocianas? Responda através dum teste de hipóteses adequado, exigindo o ónus da prova à afirmação.

(b) É possível excluir um preditor sem afectar de forma significativa a qualidade de ajustamento do modelo. Identifique, justificando, o preditor cuja exclusão menos afecta o ajustamento.

(c) Calcule o valor do coeficiente de determinação do submodelo correspondente a excluir o preditor `fenois`. Comente.

II [5 valores]

Um ensaio com a casta Malvasia, realizado em Fontanelas, visava comparar os rendimentos de 9 diferentes genótipos (designados `MV1` a `MV9`), pretendendo-se escolher um genótipo de elevado rendimento

mas que fosse sistematicamente bom em condições meteorológicas de anos diferentes. Para cada um de 3 diferentes anos (2013, 2014 e 2017) foram obtidos valores de rendimentos de cada genótipo em 5 parcelas escolhidas ao acaso. A variância dos rendimentos nas 135 parcelas foi 2.012446 (kg/planta)^2. As médias global, por ano, por genótipo e por combinação ano/genótipo, são indicadas de seguida.

```
Tables of means
Grand mean     ano                  genotipo
2.854985       A2013 A2014 A2017    MV1   MV2   MV3   MV4   MV5   MV6   MV7   MV8   MV9
               1.883 2.175 4.508    2.444 2.508 2.918 2.948 2.886 3.173 3.413 2.795 2.609

 ano:genotipo
       genotipo
ano      MV1   MV2   MV3   MV4   MV5   MV6   MV7   MV8   MV9
  A2013 1.178 1.920 1.736 2.189 2.086 2.108 2.416 1.797 1.513
  A2014 2.534 1.564 2.217 1.725 2.003 2.588 2.638 2.007 2.295
  A2017 3.621 4.041 4.800 4.931 4.570 4.822 5.186 4.580 4.020
```

1. Identifique o delineamento experimental usado e descreva em pormenor o modelo ANOVA adequado à experiência.

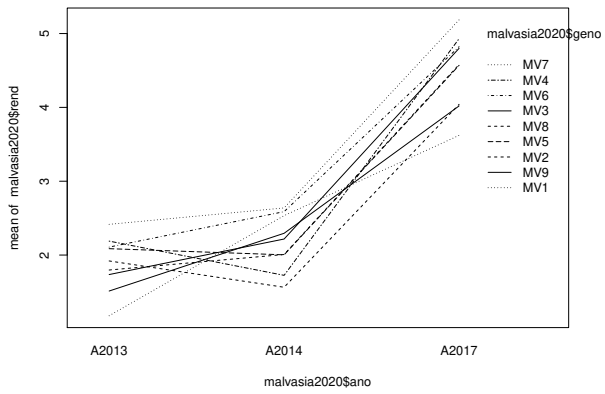
2. Complete a seguinte tabela ANOVA, indicando como obtém os valores omissos indicados pelos pontos de interrogação.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
ano	???	186.33	93.16	???	<2e-16
genotipo	???	11.69	???	2.566	---
ano:genotipo	???	10.17	0.64	1.117	0.3490
Residuals	???	???	0.57		

3. Que tipos de efeitos devem ser considerados significativos? Caso necessite de realizar mais do que um teste, descreva um em pormenor e o(s) restante(s) de forma sintética.

4. Compare os rendimentos do genótipo `MV7` nos três anos do ensaio e diga quais as diferenças que se devem considerar significativas. **Nota:** O quantil de ordem 0.95 na distribuição relevante para o teste é 5.369266.

5. Descreva e comente o seguinte gráfico.



III [2 valores]

Pretende-se determinar os mecanismos genéticos que governam duas características de grãos de milho. O endosperma pode ser açucarado (traço recessivo) ou amiláceo (dominante); e a côr da aleurona (uma proteína no endosperma) pode ser purpúrea (dominante) ou branca (recessivo). Admitindo que cada uma destas características é governada por apenas um gene, com segregação independente, seria de esperar que na segunda geração do cruzamento duma linha pura de milho com grãos de aleurona purpúrea e endosperma amiláceo com outra, também pura, de aleurona branca e endosperma açucarado, se observassem 9/16 de plantas com as duas características dominantes; 1/16 com as duas características recessivas; 3/16 com grãos de aleurona branca e endosperma amilácea; e 3/16 com grãos de aleurona purpúrea e endosperma açucarado. Realizado um cruzamento de duas linhas puras acima descritas, observaram-se na segunda geração as contagens indicadas na tabela. Teste se os dados são compatíveis com a hipótese genética referida ($\alpha = 0.05$). Comente as suas conclusões e, em caso de rejeição da hipótese nula, discuta razões dessa rejeição.

	Endosperma	
Aleurona	amiláceo (dominante)	açucarado (recessivo)
purpúrea (dominante)	248	56
branca (recessivo)	56	48

IV [4 valores]

Considere uma regressão linear múltipla da variável Y sobre p variáveis preditoras, ajustada com base em n observações.

1. Descreva o triângulo rectângulo do espaço das variáveis ($\mathbb{R}^n$) que está *directamente* relacionado com a fórmula fundamental da regressão linear. A qual conceito *geométrico* é que corresponde, nesse triângulo, a razão entre a proporção da variabilidade de Y explicada, e não explicada, pela regressão?

2. Considere agora um submodelo, com k preditores.

   (a) Mostre que o R^2 *modificado* do submodelo é maior que o do modelo completo se e só se a estimativa da variância dos erros aleatórios no submodelo for menor que no modelo completo.

   (b) Mostre que a desigualdade $QMRE_c > QMRE_s$ (onde os índices c e s indicam, respectivamente, o modelo completo e o submodelo) é equivalente a dizer que a estatística do teste F parcial comparando estes dois modelos toma valor inferior a 1.

   (c) Discuta a implicação das condições das alíneas anteriores para um algoritmo de exclusão sequencial baseado nos testes *t-Student*.