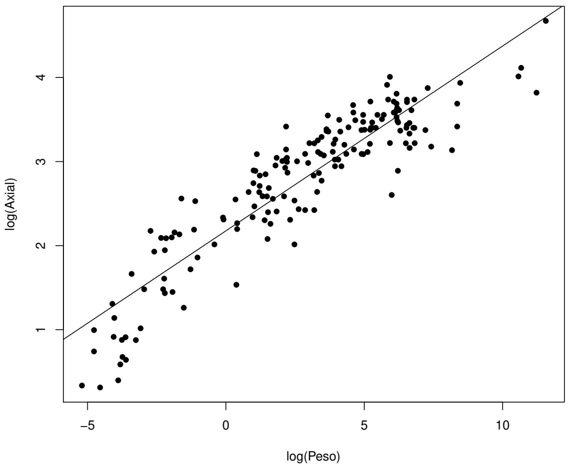
INSTITUTO SUPERIOR DE AGRONOMIA

ESTATÍSTICA E DELINEAMENTO – 2021-22

19 Janeiro 2022 **Primeira Chamada de Exame** **Duração: 3h00**

I [6 valores]

Um estudo obteve dados relativos ao peso corporal (variável `Peso`, em kg) e ao comprimento axial dos olhos (variável `Axial`, em mm) de 172 espécies de mamíferos. Após alguma exploração, foi ajustada uma regressão linear do log-comprimento axial sobre o log-peso corporal, podendo ver-se em baixo a respectiva nuvem de pontos e recta ajustada, bem como algumas médias e variâncias. Sabe-se que o declive da recta ajustada é 0.219779.



Variável	média	variância
Peso	1783.192	114429848
Axial	21.68597	213.0607
ln(Peso)	2.771838	13.2079
ln(Axial)	2.785572	0.7682822

1. Determine a equação da recta ajustada. Calcule e discuta o seu coeficiente de determinação.

2. Teste o ajustamento global do modelo. Comente, tendo também em conta o gráfico acima.

3. Qual a estimativa da variância σ^2 dos erros aleatórios do modelo? Justifique a sua resposta.

4. Deduza a relação não linear entre o comprimento axial dos olhos e o peso do corpo (não logaritmizados), correspondente à regressão linear acima ajustada. Interprete a relação obtida.

5. Construa um intervalo a 95% de confiança para o declive da recta de regressão populacional. Com base nesse intervalo, comente a seguinte observação: *"na população, a quinta potência do comprimento axial é proporcional ao peso do corpo"*. **Nota:** Caso não tenha resolvido a pergunta 3, use o valor 0.130.

II [4,5 valores]

Um estudo realizado em 2020 sobre a casta Tinta Caiada, de vinhos tintos, efectuou 219 medições independentes, em parcelas seleccionadas ao acaso, de várias características: teor brix do mosto (variável `brix`, em graus brix); volume do mosto do bago (variável `volume`, em ml/bago); peso dos bagos (variável `peso`, em g/bago); `pH` (variável com esse nome, na respectiva escala); acidez total (variável **acidez**, em g/l de ácido tartárico); teor de antocianas (variável **antocianas**, em mg/l); teor de fenóis totais (variável **fenois**, em mg/l) e rendimento (variável **rendimento**, em kg/planta). Eis alguns indicadores, bem como a matriz de correlações, relativas aos dados:

	volume	peso	pH	brix	acidez	antocianas	fenois	rendimento
mínimo	0.646	1.140	3.713	17.800	2.850	145.282	454.700	1.714
média	1.12072	1.81524	4.01786	21.78421	3.98509	507.30086	826.46711	5.67075
máximo	1.775	2.728	4.307	31.543	5.250	875.099	1335.498	12.764
desv. pad.	0.16970	0.24374	0.10167	1.71668	0.37212	114.01540	168.77408	1.80235

```
               volume     peso       pH     brix   acidez antocianas   fenois rendimento
volume      1.00000  0.97225  0.09839  0.10257  0.29780   -0.05139 -0.13216    0.01261
peso        0.97225  1.00000  0.14642  0.17528  0.27445   -0.02050 -0.10315   -0.01097
pH          0.09839  0.14642  1.00000  0.62596 -0.30396    0.27795  0.33202   -0.53662
brix        0.10257  0.17528  0.62596  1.00000 -0.18284    0.62712  0.60187   -0.57781
acidez      0.29780  0.27445 -0.30396 -0.18284  1.00000   -0.40599 -0.46341   -0.06331
antocianas -0.05139 -0.02050  0.27795  0.62712 -0.40599    1.00000  0.89770   -0.28760
fenois     -0.13216 -0.10315  0.33202  0.60187 -0.46341    0.89770  1.00000   -0.30180
rendimento  0.01261 -0.01097 -0.53662 -0.57781 -0.06331   -0.28760 -0.30180    1.00000
```

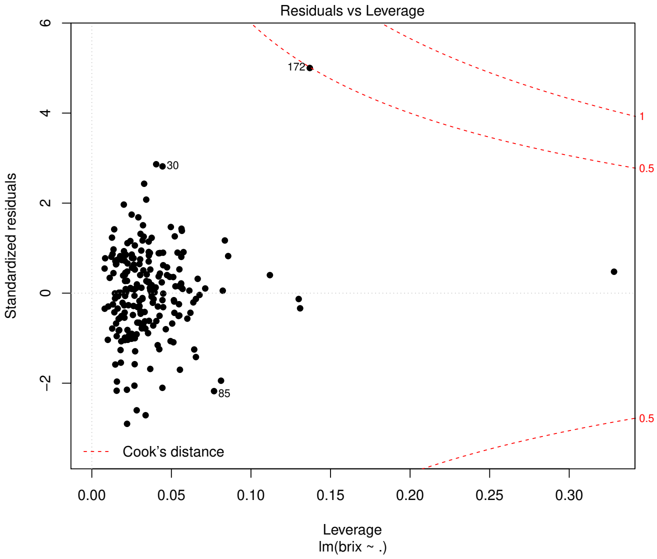
Foi ajustado um modelo de regressão linear múltipla para modelar o teor de brix a partir dos restantes preditores. Eis os resultados obtidos:

```
Call: lm(formula = brix ~ ., data = TintaCaiada)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.9555956  4.0543928  -2.209 0.028260
volume      -6.4281726  1.6617136  -3.868 0.000146
peso         5.1203644  1.1646730   4.396 1.74e-05
pH           5.9547256  0.8776247   6.785 1.15e-10
acidez       0.5038065  0.2304930   2.186 0.029932
antocianas   0.0061633  0.0013151   4.686 4.98e-06
fenois       0.0009412  0.0009197   1.023 0.307314
rendimento  -0.2095355  0.0476012  -4.402 1.70e-05
---
Residual standard error: 0.9516 on 211 degrees of freedom
Multiple R-squared: 0.7026,  Adjusted R-squared: 0.6927
F-statistic:  71.2 on 7 and 211 DF,  p-value: < 2.2e-16
```

1. É sugerido que os coeficientes dos preditores **antocianas** e **fenois** no modelo possam ser iguais, dada a natureza das variáveis e a sua medição com as mesmas unidades de medida. Efectue um teste de hipóteses adequado para avaliar esta hipótese, sabendo que as linhas e colunas da matriz de (co-)variâncias dos estimadores dos parâmetros que correspondem a estes preditores são as seguintes:

```
> vcov(TintaCaiadaBrix.lm)[c(6,7),c(6,7)]
              antocianas        fenois
antocianas  1.729520e-06 -1.043154e-06
fenois     -1.043154e-06  8.459073e-07
```

2. Descreva e discuta o seguinte gráfico. Em particular, comente o ponto que surge no topo e o ponto que surge mais à direita no gráfico.



3. Comente a seguinte afirmação: "*O melhor submodelo com apenas dois preditores é o submodelo com os preditores* `pH` *(x_1) e* `antocianas` *(x_2) e o plano de regressão resultante tem equação* $y = -8.9555956 + 5.9547256\,x_1 + 0.0061633\,x_2$".

4. O modelo de regressão linear múltipla que resulta de excluir, do modelo inicial, os preditores `acidez` e `fenois`, tem Coeficiente de Determinação $R^2 = 0.6952$. Teste se esse submodelo tem um ajustamento significativamente pior do que o modelo original.

III [5,5 valores]

Num estudo sobre selecção clonal da variedade de macieira Bravo de Esmolfe foram estudados sete genótipos, designados `B020`, `B054`, `B164`, `B206`, `B226`, `B241` e `B263`. No ensaio, foram aleatoriamente associadas parcelas a cada genótipo e mediu-se, em kg/planta, o rendimento em cada parcela, dois anos após a plantação das macieiras. Uma descrição inicial do ensaio indicou ter-se usado um delineamento experimental a um único factor, equilibrado, com cinco repetições para cada genótipo. As médias e variâncias dos rendimentos, para a totalidade das parcelas e para as parcelas associadas a cada genótipo são indicadas em baixo.

Observações	B020	B054	B164	B206	B226	B241	B263	Totalidade
Média	0.5280	1.3170	1.1186	1.0714	2.1526	1.0434	0.8172	1.14974
Variância	0.15680	0.24099	0.42720	0.24901	0.33339	1.05899	0.50835	0.57780

1. Descreva em pormenor o modelo ANOVA adequado ao problema sob estudo.
2. Construa a tabela de síntese da ANOVA, indicando como obtém cada um dos seus valores.
3. É possível rejeitar a hipótese de todos os genótipos terem o mesmo rendimento populacional médio? Justifique a sua resposta de forma pormenorizada.
4. O genótipo com o maior rendimento médio amostral é o genótipo `B226`. É possível afirmar que o respectivo rendimento amostral difere significativamente do de qualquer outro genótipo?

5. Chegou entretanto a informação que a descrição do delineamento não tinha sido feita de forma correcta, uma vez que o ensaio foi realizado em cinco diferentes localidades, tendo em cada localidade sido associada (de forma aleatória) uma parcela a cada um dos sete genótipos.

   (a) Identifique o delineamento experimental que tinha, afinal, sido utilizado no ensaio. Escreva a equação do modelo ANOVA correspondente.

   (b) Sabendo que a Soma de Quadrados associada às diferentes localidades é 3.104, construa a nova tabela-resumo da ANOVA.

   (c) Diga, de forma sintética, se a sua resposta à pergunta 3 mudaria face à nova informação recebida. Comente.

IV [4 valores]

1. Considere os indicadores R^2 usual e modificado numa regressão linear múltipla com p preditores e ajustado com base em n observações.

   (a) Mostre que se verifica a relação $R^2_{mod} = \frac{-p}{n-(p+1)} + R^2 \frac{n-1}{n-(p+1)}$.

   (b) Indique, justificando, a gama de valores que R^2_{mod} pode tomar.

2. Considere uma regressão linear múltipla com p preditores e ajustada com n observações.

   (a) Escreva o modelo em notação vectorial/matricial, identificando as quantidades utilizadas.

   (b) Mostre que o vector dos valores ajustados $\hat{Y}_i$ se pode escrever como $\vec{\hat{\mathbf{Y}}} = \mathbf{X}\vec{\boldsymbol{\beta}} + \mathbf{H}\vec{\boldsymbol{\epsilon}}$, sendo $\mathbf{H}$ a matrix de projecção ortogonal sobre o espaço das colunas da matriz do modelo.

   (c) A partir da alínea anterior,

      i. mostre que a matriz de variâncias-covariâncias do vector $\vec{\hat{\mathbf{Y}}}$ é dada por $\sigma^2\mathbf{H}$;

      ii. relacione a variância de cada valor ajustado $\hat{Y}_i$ com o respectivo efeito alavanca;

      iii. mostre que, dado o modelo, a variância de $\hat{Y}_i$ não pode exceder a variância de Y_i.